The cover image was generated with GPT-5

INAUGURAL-DISSERTATION

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Tim Raedsch, M. Sc.

aus Schwabach (Nuernberg)

Tag der mündlichen Prüfung : 21.01.2026

From Annotation to Validation: Strengthening Machine Learning Pipelines for Real-World Robustness

Betreuerin: Prof. Dr. Lena Maier-Hein
Zweitbetreuerin: Prof. Dr. Annette Kopp-Schneider

Abstract

Artificial Intelligence (AI) is transforming science, industry, and daily life at an unprecedented pace. Foundation Models (FMs) now shape how we search, create, and interact with information, across domains from finance to healthcare. Yet, paradoxically, most AI research produced in academic and industrial labs remains confined to benchmark performance and paper metrics. The rapid increase in computational resources and in the size of the research community has contributed to this progress by fueling an unprecedented surge in publications, with venues such as Conference on Neural Information Processing Systems (NeurIPS) exhibiting an annual growth rate of approximately 29% since 2017. This apparent progress is reflected in the continual stream of reported state-of-the-art (SOTA) results for the majority of publications. Yet, despite the vast number of proposed methods and the accompanying media attention, only a small fraction translate into real-world impact. This discrepancy raises a fundamental question: how can methodological research better align with the goal of practical translation?

This thesis addresses overlooked elements of the model development pipeline that critically influence whether research outputs achieve translation. We investigate tasks situated upstream of model training, such as labeling instructions and internal Quality Assurance (QA), as well as downstream tasks, including user-centric benchmarking. These questions are particularly urgent in healthcare, where methodological weaknesses can translate into risks for patient safety. To illustrate this challenge, we focus on surgery as a representative high-stakes domain. Across these studies, we demonstrate that design choices in these stages can and often exert a greater impact on model performance and applicability than many architectural innovations. Building on our analysis, we propose new methods and best practices for sustainable research and reliable translation.

First, we demonstrate that the majority of biomedical imaging projects does not provide any labeling instructions, even though annotators consistently consider them crucial. In the largest study on labeling instructions to date to our knowledge, we demonstrate that visual examples are key to improving annotation quality, whereas solely extending textual descriptions yields no measurable or even negative benefit. To address these shortcomings, we provide best practices for involving annotators in both the annotation process and the construction of labeling instructions. Following a comprehensive review by the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) Special Interest Group for Biomedical Image Analysis Challenges [119], future challenges are now required to adhere to stricter standards, including the mandatory publication of labeling instructions.

Second, we turn our focus to internal QA, which at the time represented a standard practice in annotation workflows for professional annotation companies. We demonstrate that this additional step yields only marginal improvements in annotation quality. In resource-constrained scenarios, we further show that investment in labeling instructions provides greater benefits than additional QA. To guide more effective QA, we developed a statistical model and identified generalizable image characteristics for which QA is most beneficial. This ultimately enables users to selectively apply QA to images with a high likelihood of improvement, instead of processing every image.

Third, we show that with the growing number of benchmarks, domain users must carefully select their Vision-Language Model (VLM) of choice. To support users with their domain-specific benchmarking, we developed DomainBench, a framework for turning user selected image datasets

into diverse and scalable VLM benchmarks. To help the community adopt our framework, we released seven new datasets covering a wide range of domains from kitchen environments to animal scenes. To complement the generation process, we introduced a new metric, Accuracy%(t), that accounts for shared base images across different tasks, which previous work neglects.

Lastly, we apply DomainBench to the surgical domain, where robust translation of AI methods is especially critical, and provide the first comprehensive assessment of VLMs in surgical imaging by benchmarking 15 SOTA VLMs across more than 150,000 question-answer pairs. We demonstrate that current VLMs can handle elementary surgical perception tasks, but struggle on questions that require medically informed reasoning. We further provide a critical comparison of medical versus generalist VLMs, uncovering that specialized medical VLMs are outperformed by generalist VLMs. Our findings emphasize the need for future work on knowledge integration strategies that extend beyond conventional finetuning.

Together, these contributions highlight that translation in Machine Learning (ML) depends less on incremental architectural advances than on careful design the model development pipeline. By addressing labeling instructions, internal QA, user-centric benchmarking, and domain-specific validation, this thesis provides both methodological insights and practical frameworks that strengthen the reliability and applicability of ML Models. In doing so, it contributes to aligning methodological progress with real-world impact across natural and medical imaging, and ultimately improving the scientific standards in the field.

Zusammenfassung

Künstliche Intelligenz (KI) verändert Wissenschaft, Industrie und Alltag in beispiellosem Tempo. FMs prägen mittlerweile, wie wir nach Informationen suchen, Inhalte erstellen und mit Wissen interagieren in allen Bereichen von der Finanzwelt bis zur Gesundheitsforschung. Paradoxerweise bleibt jedoch der Großteil der in akademischen und industriellen Forschungsabteilungen erzeugten KI-Forschung auf Benchmark-Leistungen und Papiermetriken beschränkt. Der rapide Anstieg an Rechenressourcen und die wachsende Größe der Forschungsgemeinschaft haben zu einer beispiellosen Zunahme an Publikationen geführt. Beispielsweise verzeichnen Konferenzen wie NeurIPS seit 2017 eine jährliche Wachstumsrate von rund 29%. Dieser scheinbare Fortschritt spiegelt sich in der kontinuierlichen Veröffentlichung neuer SOTA-Ergebnisse wider. Dennoch gelangt nur ein kleiner Teil der vorgeschlagenen Methoden und der damit einhergehenden medialen Aufmerksamkeit tatsächlich in die Praxis. Diese Diskrepanz wirft eine grundlegende Frage auf: Wie kann die methodische Forschung stärker mit dem Ziel einer praktischen Umsetzung in Einklang gebracht werden?

Diese Dissertation widmet sich bislang vernachlässigten Elementen des Modellentwicklungsprozesses, die maßgeblich bestimmen, ob Forschungsergebnisse in die Praxis übertragen werden. Untersucht werden Schritte im Vorfeld des Modelltrainings, wie Annotationsanleitungen und interne Qualitätssicherung, sowie nachgelagerte Schritte, darunter nutzerzentriertes Benchmarking. Diese Fragestellungen sind insbesondere im Gesundheitswesen von großer Bedeutung, da methodische Schwächen hier unmittelbare Auswirkungen auf die Patientensicherheit haben können. Um diese Herausforderung exemplarisch zu beleuchten, konzentrieren wir uns auf die Chirurgie als ein repräsentatives Hochrisikofeld. Über diese Studien hinweg zeigen wir, dass Gestaltungsentscheidungen in diesen Phasen oftmals stärkeren Einfluss auf Modellleistung und Anwendbarkeit haben als viele Modifikationen der Modellarchitektur. Aufbauend auf dieser Analyse entwickeln wir neue Methoden und Best Practices für nachhaltige Forschung und verlässliche Translation.

Erstens zeigen wir, dass die Mehrheit biomedizinischer Bildgebungsprojekte keine Annotationsanleitungen bereitstellt, obwohl Annotatoren diese durchweg als unverzichtbar ansehen. In der bislang größten Studie zu Annotationsanleitungen weisen wir nach, dass visuelle Beispiele entscheidend für eine Verbesserung der Annotationsqualität sind, während die bloße Erweiterung der Textbeschreibungen keinen oder sogar einen negativen Effekt hat. Zur Behebung dieser Nachteile schlagen wir Best Practices vor, die Annotatoren sowohl in den Annotationsprozess als auch in die Erstellung von Richtlinien mit einbeziehen. Nach einer umfassenden Überprüfung durch die MICCAI Special Interest Group for Biomedical Image Analysis Challenges [119] sind künftige Wettbewerbe verpflichtet, strengere Standards einzuhalten, einschließlich der verpflichtenden Veröffentlichung von Annotationsanleitungen.

Zweitens richten wir den Fokus auf interne Qualitätssicherung, die bereits damals eine Standardpraxis in Annotations-Workflows professioneller Anbieter darstellte. Wir zeigen, dass dieser zusätzliche Schritt die Annotationsqualität nur marginal verbessert. In ressourcenbeschränkten Szenarien belegen wir zudem, dass Investitionen in Annotationsanleitungen größeren Nutzen bringen als zusätzliche Qualitätssicherung. Um Qualitätssicherung gezielter einzusetzen, entwickelten wir ein statistisches Modell und identifizierten verallgemeinerbare Bildcharakteristika, bei denen Qualitätssicherung besonders effektiv ist. Dies erlaubt es, Qualitätssicherung selektiv auf Bilder mit hoher Verbesserungswahrscheinlichkeit anzuwenden, anstatt alle Bilder gleichermaßen zu prüfen.

Drittens zeigen wir, dass angesichts der wachsenden Zahl an Benchmarks Domänennutzer ihre Wahl eines VLMs mit Bedacht treffen müssen. Zur Unterstützung nutzerspezifischer Evaluation entwickelten wir DomainBench, ein Framework, das benutzerdefinierte Bilddatensätze in vielfältige und skalierbare VLM-Benchmarks überführt. Um die Anwendung zu erleichtern, veröffentlichen wir sieben neue Datensätze, die ein breites Spektrum von Küchenumgebungen bis hin zu Tierszenen abdecken. Ergänzend führen wir eine neue Metrik, Accuracy

Schließlich wenden wir DomainBench im chirurgischen Bereich an, wo eine robuste Translation von KI-Methoden besonders kritisch ist, und präsentieren die erste umfassende Auswertung von VLMs in der chirurgischen Bildgebung. Dabei benchmarkten wir 15 SOTA-Modelle anhand von über 150.000 Frage-Antwort-Paaren. Wir zeigen, dass aktuelle VLMs grundlegende chirurgische Perzeptionsaufgaben bewältigen können, jedoch bei Fragen scheitern, die medizinisch fundiertes Wissen erfordern. Darüber hinaus liefern wir einen kritischen Vergleich zwischen medizinischen und generalistischen VLMs und zeigen, dass spezialisierte medizinische Modelle von generalistischen VLMs übertroffen werden. Unsere Ergebnisse verdeutlichen die Notwendigkeit künftiger Forschungsarbeiten zu Strategien des Wissenstransfers, die über herkömmliches Finetuning hinausgehen.

Zusammenfassend verdeutlichen diese Beiträge, dass Translation in der ML-Forschung weniger von inkrementellen Architekturfortschritten abhängt als von der sorgfältigen Gestaltung des Modellentwicklungsprozesses. Durch die Berücksichtigung von Annotationsanleitungen, interner Qualitätssicherung, nutzerzentriertem Benchmarking und domänenspezifischer Validierung liefert diese Arbeit sowohl methodische Einsichten als auch praktische Werkzeuge, die die Zuverlässigkeit und Anwendbarkeit von ML-Modellen stärken. Damit trägt sie dazu bei, methodischen Fortschritt enger mit Translation in Natur- und medizinischer Bildgebung zu verbinden und die wissenschaftlichen Standards des Feldes nachhaltig zu verbessern.

Acknowledgements

There are many individuals I wish to thank for shaping my experience during these four years as a doctoral student in Heidelberg.

I am deeply grateful to my adviser, Lena Maier-Hein, whose guidance through countless meetings and exchanges transformed me into a more confident researcher. Their passion, vision, and commitment to doing science that is sustainable and meaningful have been a constant source of motivation. They always emphasized quality over quantity and bet on research directions that were uncommon at the time, such as benchmarking, which has since become much more widely recognized. From them I learned not only technical knowledge and the art of communicating ideas, but also how to think carefully about what constitutes good and lasting science.

Moreover, I would like to express my deep gratitude to my second supervisor, Prof. Dr. Annette Kopp-Schneider, whose statistical expertise proved invaluable throughout this work. Her advice and feedback significantly shaped my projects, and I am especially thankful for the emphasis she placed on the necessity of rigorous statistical analysis. This is regrettably not always standard in current scientific practice and not appreciated enough.

Many thanks to my peers in the Validation of Intelligent Systems (VIS) subgroup: Annika Reinke, Matthias Eisenmann, A. Emre Kavur, and Evangelia Christodoulou. It has been a privilege to work with such a supportive and kind group, always ready to lend a hand or share a good conversation. I owe special thanks to Matthias, who accompanied me as an office mate throughout the Covid time in the first two years. Furthermore I owe special thanks to Annika for her great support as a teamlead and her exceptional illustration skills. Emre deserves special mention for his tireless motivation and his constant drive to always go for another round of rage cage.

Financial support from the Hanns Seidel Foundation scholarship with funds from the Federal Ministry of Education and Research Germany (BMBF) is gratefully acknowledged. The scholars are a remarkable group of people that contribute both intellectually and voluntarily to the community.

Daily life during my PhD was made immeasurably richer by my colleagues and friends in SYMIC and DKFZ. The camaraderie, discussions, and support within this community created an atmosphere where both personal and professional growth could flourish. I would like to thank, in alphabetical order, Michael Baumgartner, Markus Bujotzek, Till Bungert, Simone Menghai Clas, Kris Dreher, Minu Dietlinde Tizabi, Nico Disch, Stefan Denner, Helena Engel, Lorenz Feineis, Ralf Floca, Karol Gotkowski, Martin Grözinger, Benjamin Hamm, Doreen Heckmann-Nötze, Chris Heidt, Niklas Holzwarth, Elisa Holstein, Fabian Isensee, Piotr Kalinowski, Yannick Kirchhoff, Lukas Klein, Robert Knörlein, Marcel Knopp, Balint Kovacs, Lars Kraemer, Vera Klütz, Carsten Lueth, Leon Mayer, Keno März, Marco Nolden, Jan-Hinrich Nölke, Tobias Norajitra, Marco Pascale, Ahmad Bin Qasim, Maike Rees, Saikat Roy, Tom Rix, Jan Sellner, Silvia Seidlitz, Max Schwab, Thuy Nuong Tran, Tassilo Wald, Daniel Walther, Tim Robin Wagner, Amine Yamlahi, Sebastian Ziegler, David Zimmerer, and Ferdinand Zettl for their companionship, stimulating conversations, and many memorable moments.

I am equally grateful to the office with Nina Kraft, Michaela Gelz, Theresa Klocke and Stefanie Strzysch for the wonderful atmosphere and support. You are always invited for a Lebkuchen Lolli-Pop in Nuremberg, provided that they are available.

Lastly i would like to thank my family, who have supported me throughout my life and who

have always been proud of me and believed in me. I would also like to thank my friends Max Kippenberg and Christian Maißler for always being there whenever I needed someone to talk to, and for all the fun times we shared.

1 Introduction

This chapter first motivates the need for reliable evaluation in modern Machine Learning (ML) and situates it in the broader context of Foundation Models (FMs) and Vision-Language Models (VLMs) (Section 1.1). Next, Section 1.2 formulates the four research questions that guide this thesis, focusing on how annotation practices and benchmark design affect the reliability of evaluations. Finally, Section 1.2.2 provides an outline of the thesis structure to facilitate navigation for the reader.

1.1 Motivation

> *This is because good evals are very difficult to build - at Tesla I probably spent 1/3 of my time on data, 1/3 on evals, and 1/3 on everything else.*
>
> — Andrej Karpathy

ML has seen rapid growth over the past decade, with the number of publications in the field rising dramatically year by year [9]. This fast-paced development has brought major changes in how ML research is conducted and applied, but it has also created new challenges that remain unresolved.

One of the most important shifts is the emergence of so-called FMs, which are large models trained on broad datasets and then adapted to solve a wide variety of tasks [140]. In contrast to traditional ML, where a model was typically designed for a specific, well-defined problem, FMs are intended to be flexible and general. The goal is to reduce the need for task-specific engineering and allow a single model to serve as a starting point for many different applications. Recent advances from OpenAI, Anthropic, and Google reflect this trend: Gemini 2.0, for instance, can perform image editing tasks directly from textual prompts [85].

At the same time, concerns about reproducibility have become more prominent [118, 29, 64]. Many findings in the literature cannot be reproduced reliably [86], and the practical impact of published research remains limited in many areas [118]. Problems such as hidden data processing steps [155], poorly documented methods [56], and changing experimental setups [168] all contribute to this situation. These issues become even more critical when working with large, complex FMs, where

even small inconsistencies in data, instructions, or evaluation setup can lead to large shifts in reported results.

To ensure robust validation in the age of FMs, we have to go beyond model architectures or training pipelines. Instead, we turn our attention to two often overlooked but deeply interlinked elements: annotation quality and benchmark design. While these two aspects are often treated independently, they are fundamentally connected: reliable annotations form the basis of any benchmark, and the structure of a benchmark in turn shapes what annotation quality is required. Together, they define the foundation by which model performance is interpreted. For instance, in biomedical imaging, the same model may appear to 'fail' or 'succeed' depending on how rigorously the labeling instruction was enforced, and how diverse the benchmark scenarios are. Our thesis shows how variation in either element, annotation or benchmark, can lead to misleading conclusions about model capability.

First, we investigate the role of labeling instructions and Quality Assurance (QA) in shaping annotation quality. These elements are frequently underreported or neglected in both academic and industrial practice [168], despite being critical for ensuring that the data used for training and validation is meaningful and consistent. Our results show that the way annotators are instructed and how their work is reviewed, has a substantial effect on the resulting annotations and, by extension, on the models trained or evaluated on those annotations [155].

Second, we develop a framework for turning user selected image datasets into diverse and scalable VLM benchmarks. The key idea is to construct task variants using human-curated, heuristic, and model-derived metadata, without relying on prompt-only generation. This enables users to transform their own domain-specific data into targeted benchmarks that are both cost-effective and robust. Compared to existing approaches, our method reduces the risk of hallucinated task definitions and preserves the structure of the original vision tasks.

Our goal is to help close the gap between rapid technical progress in ML and the need for dependable, real-world solutions, especially in areas where reliability and robustness are non-negotiable, such as biomedical imaging. By addressing both the upstream side of annotation and the downstream side of benchmarking, this thesis aims to contribute to a more systematic understanding of what reliable validation means by quantifying the effects of annotation noise and proposing a scalable, metadata-driven benchmarking framework.

Disclosure

The research documented in this thesis represents collaborative work undertaken jointly with multiple coauthors. Accordingly, the collective pronoun "we" will be employed throughout the text to reflect the shared nature of the contributions. In every section of the contribution Chapter 4, a disclosure states my contributions and the team members of the presented work. Generative AI was used to support the formulation of this document. Images or image components generated with AI are explicitly marked, while licenses for incorporated external images are indicated alongside each figure.

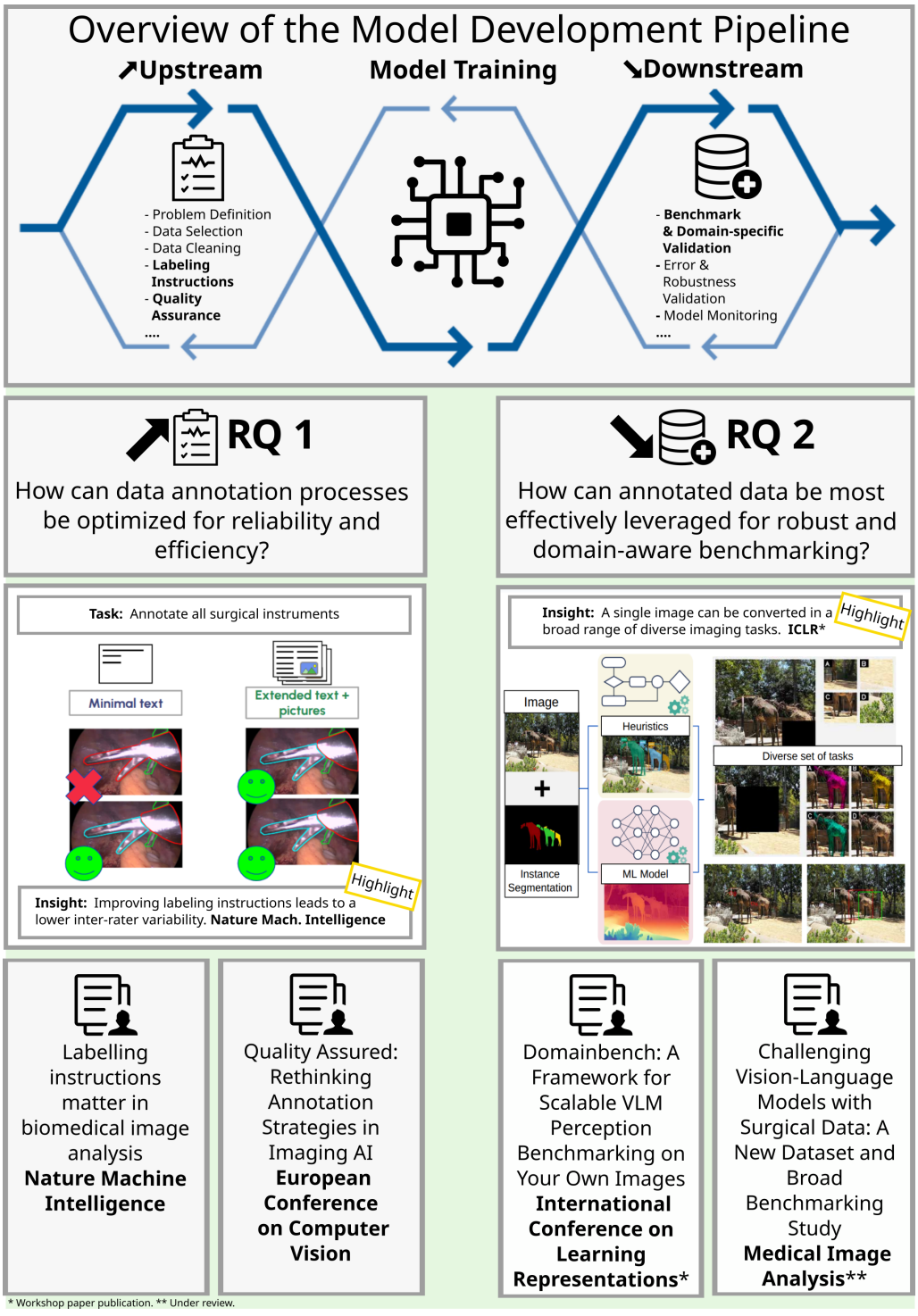


Figure 1.1: **The model development pipeline and focal research questions.** The model development pipeline is divided into three stages: upstream, model training, and downstream. Upstream involves defining the problem, selecting and cleaning data, generating labeling instructions and quality assurance for reliable annotations among others. Downstream addresses benchmarking and domain-specific validation, emphasizing robust, context-aware evaluation and error analysis. **RQ1** investigates how annotation processes can be optimized for reliability and efficiency. **RQ2** explores how annotated data can be leveraged for domain-aware benchmarking. Together, these studies contribute frameworks for improved data reliability and robust, domain-specific validation across the model lifecycle.

1.2 Research Questions

The objective of this thesis is to investigate how critical yet often under-specified design choices in data annotation and model validation affect the reliability of ML systems in biomedical imaging and beyond. To this end, we examine both upstream practices involved in constructing annotated datasets and downstream practices for validating model capabilities on domain-specific tasks.

The work is guided by two Research Questions (RQs). The first RQ addresses the upstream stage of the model development pipeline and focus on annotation quality, analyzing the impact of labeling instructions, and QA procedures on the consistency and accuracy of human-generated labels. The latter RQ address the downstream stage with the challenge of validating VLMs on user-defined datasets, and with a particular emphasis on the distinct requirements posed by surgical image data. Together, these questions aim to provide a principled basis for improving both the robustness of reference data and the credibility of model validation.

1.2.1 RQ1: How can data annotation processes be optimized for reliability and efficiency?

Human annotation remains the cornerstone of reference test sets in biomedical ML. The quality of these annotations not only determines the reliability of performance metrics but also forms the basis for any meaningful comparison between models. Despite this foundational role, most reported differences between ML methods in medical imaging are marginal and frequently fail to hold up under rigorous scientific scrutiny [29]. While recent improvements in dataset documentation, such as datasheets for datasets [56], have addressed some gaps in transparency, large segments of the data generation process, including annotation protocols, remain insufficiently described [168]. Every stage of dataset construction involves a series of design choices that can influence annotation quality, ranging from who performs the annotations and what data are selected, to how labeling instructions are formulated and which tools are used. Such decisions collectively impact the trustworthiness of reference datasets and, by extension, the reproducibility and interpretability of downstream benchmarking and clinical deployment.

To address the persistent uncertainty surrounding the annotation process, we designed a large-scale study to isolate the effects of labeling instructions on annotation quality. Specifically, we recruited over 800 annotators from both professional annotation companies and crowdsourcing platforms, instructing them to annotate more than 14,000 biomedical images. The instruction sets provided to annotators systematically varied in their level of detail and inclusion of example images. This experimental design aimed to clarify how the structure and specificity of labeling instructions alone may influence the accuracy and consistency of resulting annotations, independent of other confounders. An overview is shown in Figure 4.1.

Building on our analysis of labeling instructions, we now turn to QA as a further mechanism for safeguarding the integrity of reference annotations in the upstream stage. In this context, QA encompasses a set of systematic processes aimed at reviewing, validating, and where necessary, correcting annotations after their initial creation [144]. Such post-hoc validation steps are widely regarded as critical for improving annotation quality, particularly in domains such as medical imaging that are characterized by substantial inter-rater variability [141]. Here, inter-rater variability [83] denotes the degree of disagreement observed when multiple annotators independently label the same data point; this variability can arise from ambiguous data, inconsistent instruction,

or diverging levels of annotator expertise, and it remains a persistent source of uncertainty in both research and clinical workflows.

Despite the recognized importance of QA, it is well understood that these practices are both time- and resource-intensive [83]. The field has developed a broad array of QA strategies [144], ranging from automated error detection and expert adjudication to multi-stage consensus (see Section 3.1 for an overview). Yet, the specific contribution of internal QA, such as annotation company review pipelines, to overall annotation quality has not been systematically quantified, nor have its benefits been weighed against its substantial operational costs.

To address this gap, our study isolates the impact of internal QA by comparing annotation outcomes before and after QA processes have been applied, as schematically depicted in Figure 4.7. Leveraging a large-scale dataset comprising 57,648 laparoscopic instance masks produced by multiple annotation providers and a crowdsourcing platform, we assess both the magnitude and context-dependence of quality improvements attributable to internal QA measures. Our design enables a focused validation of whether, and under what circumstances, QA practices yield meaningful gains in annotation accuracy, providing an empirical basis for optimizing annotation pipelines in future biomedical image analysis projects.

1.2.2 RQ2: How can annotated data be most effectively leveraged for robust and domain-aware benchmarking?

With the increasing prominence of FMs, and VLMs in particular, the established paradigm for model validation has shifted. While traditional deep learning architectures were benchmarked predominantly on narrowly defined, problem-specific datasets, VLMs are now subjected to a broad spectrum of tasks intended to capture their intended generality [140]. This diversification of benchmarks spans a broad task range, including object detection, segmentation, captioning, and Visual Question Answering (VQA) among others, reflecting to comprehensively map the diverse capabilities of these models [214]. However, this general-purpose benchmarking often fails to coincide with the actual, domain-specific requirements of many researchers, thereby engendering a persistent evaluative gap [103].

Researchers are seldom interested in performance aggregates across disjointed, generic benchmarks. Instead, they are invested in understanding how such models perform on tasks that precisely reflect their own data characteristics and constraints [103]. This misalignment poses significant methodological challenges, because it is both inefficient and potentially misleading to infer domain suitability from broad, decontextualized validations.

As established in RQ1, the foundation for any meaningful assessment of VLMs lies in the meticulous curation of annotated datasets. We assume, henceforth, the availability of such high-fidelity data for the validation process.

The evaluative gap encountered by domain researchers stems less from a deficiency of raw data than from the inadequacy of existing benchmarks to capture the diversity and specificity of real-world scientific datasets. In many specialized fields, canonical vision-language benchmarks fail to encompass the phenomena or annotation regimes that are relevant to the user's domain [60].

To address these limitations, we present DomainBench, a scalable framework that systematically maps a wide range of established vision-language perception tasks onto annotated images provided by the user. Leveraging image metadata, heuristics, and minimal additional input, the

framework automates task instantiation at scale and ultimately removing manual engineering efforts.

Whereas the preceding work addressed the validation of VLMs across general imaging domains, we now shift our attention to the surgical domain, which introduces a distinct set of challenges and requirements [117] as the second part of RQ2. Surgical images are characterized by their high degree of visual complexity, variable lighting, and presence of rare anatomical or pathological features. These characteristics distinguish surgical data from more conventional benchmarks and create a demanding environment for both general-purpose and specialized models.

To systematically investigate this domain shift, we applied our benchmarking framework not only to general imaging datasets but also to a newly curated surgical dataset. In this context, we assessed the performance of both large, generalist VLMs and domain-specific fine-tuned models. Our approach allows for direct comparison between model families, providing insights into the extent to which current general-purpose architectures can accommodate the unique requirements of surgical image understanding, as well as the potential benefits of domain adaptation.

Contents

2 | Fundamentals

In this chapter we describe the fundamental concepts referenced in this doctoral thesis. Section 2.1 introduces the core terminology, workflows, and Quality Assurance (QA) strategies for *Data Annotation*. Section 2.2 then outlines the principles of *validation and benchmarking*, describing common annotation types, key benchmark components, and design considerations. Building on these foundations, Section 2.3 revisits essential ideas in Deep Learning (DL), including data splits, optimization routines, and loss functions, while Section 2.4 traces the evolution of classic convolutional architectures from AlexNet to recent architectures. Finally, Section 2.5 surveys recent advances in Vision-Language Models (VLMs), highlighting the transition from contrastive pretraining to modern Foundation Model (FM) paradigms.

2.1 Data Annotation

2.1.1 Data Annotation Definition

Data annotation refers to the process of systematically assigning labels or metadata to raw data, encompassing diverse sensor outputs such as images, text, or video [155]. In this work, our primary focus is data annotation specifically for image data, including vision-language datasets. These annotations are particularly important for supervised learning, where models are trained using labeled datasets, thus explicitly providing input-output relationships. Conversely, unsupervised learning leverages unlabeled datasets to uncover inherent patterns or structures within the provided data. Semi-supervised learning combines aspects of both, utilizing a mixture of labeled, unlabeled, or weakly labeled data to improve model performance [128].

Data annotation methodologies vary, ranging from fully human-based processes and model-assisted human annotations to entirely automated systems. Historically, supervised training methods dominated, making data annotation critically important during the model training phase [155]. Poor-quality annotations during supervised training adversely affect model performance, because the added noise can impair the model's ability to learn generalizable features and overfit on spurious patterns [141]. However, over time, the prominence of supervised training has declined, reducing the significance of extensive data annotation specifically at the training stage. Today, VLMs are primarily trained via a pipeline consisting of pretraining, fine-tuning and instruction-tuning, further diminishing the role of extensive annotations during training.

Assigned label	Cherry	Airplane	Red Panda	Tomato
Segmentation Quality	Medium	Poor	Good	Good
Actual label	Apple	Airplane	Meerkat	Electronic Component
Estimated Category Quality Rank	3rd	1th	37th	1th

Figure 2.1: **Annotation errors are present in most common datasets**. Reference annotations form the basis for benchmarking different methods against each other. While most benchmarks claim to contain only high-quality annotations, in practice annotation errors are widespread, falsify model selection [135], and occur in most datasets. The chosen examples represent a general computer vision scenario, where an objective ground truth can be obtained. The Open Images V7 dataset provides quality estimates for each category. Even the highest-rated annotations in the Airplane and Tomato categories contain errors, indicating that little to no human review of the final annotations was performed. The images are extracted from the Google Open Images Dataset V7 [96].

Nonetheless, high-quality data annotations remain indispensable, particularly for benchmarking purposes using test sets [155, 135, 161]. Accurate and reliable annotations are crucial as they ensure precise, generalizable, and dependable performance measurements, maintaining their significance as a fundamental component in the validation of modern VLMs. Annotations have become even more relevant for contemporary benchmarks, as they highlight specific shortcomings in model performance (e.g. [60, 146]). As models are increasingly adopted by users, understanding precisely where models excel and where they fail becomes essential [69].

2.1.2 General Annotation Process

The general annotation process typically involves three phases: the Preparation Phase, the Annotation Phase, and the QA Phase. Each phase plays a role in shaping the consistency and reliability of the resulting dataset, although implementation details can vary considerably depending on the task, available resources, and domain, as depicted in Figure 2.2.

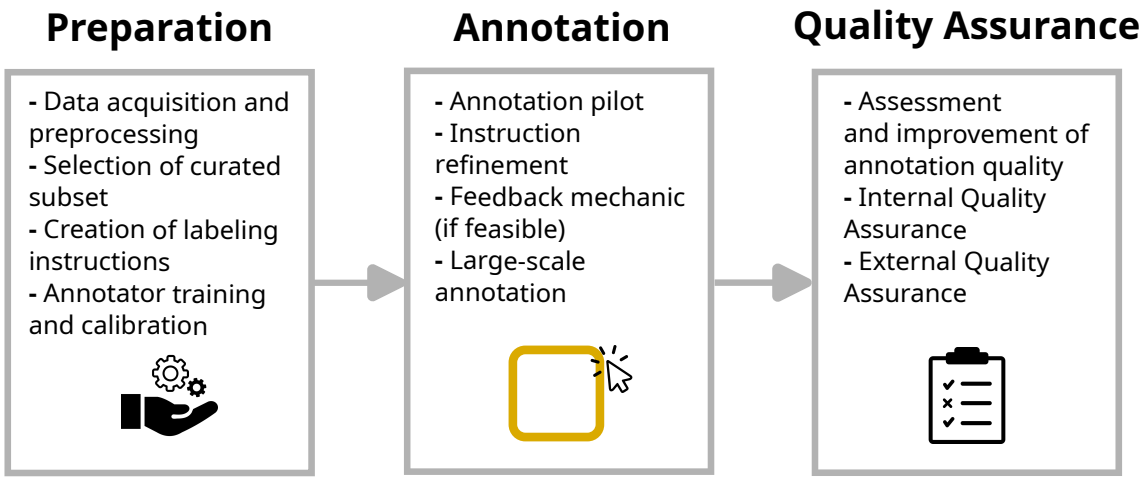


Figure 2.2: **General annotation process.** The process typically consists of a *preparation phase* (data selection, labeling instructions, training), an *annotation phase* (pilot, instruction refinement, large-scale annotation), and a *QA phase* (internal and/or external assurance to assess and improve the annotation quality).

Preparation Phase

In the preparation phase, data is acquired and preprocessed with the specific objectives of the annotation task in mind [51]. The data should be representative of the target domain and use case. Low-quality samples, such as those affected by artifacts or misclassified examples, are typically excluded. Since full annotation of all available data is often infeasible, a curated subset is typically selected [83]. This selection process may involve undersampling frequent patterns or highly repetitive cases and enriching the dataset with rare or underrepresented examples to support downstream model performance or validation needs.

After selecting the data, labeling instructions should be created to standardize the labeling process. These typically outline what elements should be annotated, how annotations should be carried out, and how to use the annotation tool. While comprehensive instructions can improve consistency, their clarity and level of detail often depend on the resources and time available for task design [51]. Of note, labeling instructions are often neglected and not even reported [155].

Annotator training and calibration are generally recommended to align annotators' understanding of the labeling instructions, particularly when subjective decisions are involved [155]. However, in many practical settings, especially in large-scale or time-constrained projects, such training is minimal or absent, potentially leading to variability in annotations.

Annotation Phase

The annotation phase involves the actual labeling of data. In some cases, teams begin with a small pilot [176] to check whether annotations align with task expectations and whether annotators interpret the instructions consistently. This step can be useful for identifying ambiguous cases or refining instructions. Where possible, feedback mechanisms are introduced to iteratively improve annotations, though their implementation depends on annotator availability and project

timelines [46]. Once a reasonable level of consistency or agreement is observed, annotation may proceed at a larger scale. For a given image, annotations can be generated by a single person or multiple people.

Quality Assurance Phase

The main goal of QA is to evaluate and where necessary improve the reliability, accuracy, and usability of generated annotations. Therefore, the QA phase focuses on assessing and improving annotation quality through review or post-hoc validation. Several approaches may be employed, including majority voting among annotators, review by domain experts or more experienced annotators, and the use of automated heuristics or models to flag data points [33]. A broad overview of internal QA and external QA is provided in the next subsection. Even in settings with automated or semi-automated annotations, human reviewers may audit a subset of the data to detect and fix systematic errors or quality issues [107]. While such review does not guarantee correctness of the annotations, it provides a level of confidence in the overall dataset quality. Of note, this confidence can often be misleading. Just because a data point has passed through several annotators does not automatically imply that the quality is sufficient. Similarly, high confidence from a model could originate from spurious correlations in the training data or calibration errors [59], rather than a true indication of label correctness.

2.1.3 Quality Assurance Approaches

QA in annotation pipelines can be broadly categorized into two types depending on the annotation exposure to the annotation requester: internal and external QA [155].

Internal QA refers to processes integrated after the initial annotation, normally by the same team conducting the labeling at the same annotation company. The requester of the annotations does not have access to the initial annotations.

External QA involves retrospective checks, typically by independent reviewers, team leads, or automated systems, to detect and correct errors after annotation completion. Thus, the requester of the annotations usually has access to the initial annotations. It should be noted that annotation processes vary between companies and change rapidly given the recent model developments.

As highlighted in the previous subsection, QA methods vary in cost and coverage and are typically selected based on task complexity and resource constraints [33]. Several QA methods are used to assess the reliability and consistency of annotations. These methods can be applied independently or in combination, depending on the task and available resources. For an in-depth examination of QA strategies, see Section 3.1.

2.2 Validation and Benchmarks

2.2.1 Annotation Types and Tasks

Before we describe benchmarks, we need to define the most common annotation types and tasks. Annotations can be assigned at varying levels of granularity and applied to different regions

within an image, depending on the intended use or task [118].

Image-level Annotations At the image level, annotations are primarily employed for classification tasks, involving tagging the entire image I with relevant labels. Typically, these annotations consist of categorical labels, such as "dog," "cat," or "heart." With the recent emergence of VLMs, image-level annotations have expanded to include richer forms of metadata like textual descriptions or question-answer pairs, facilitating tasks like image captioning and Visual Question Answering (VQA). Formally, these annotations map an image I to a label c from a predefined set of categories $\mathcal{C}$:

$$I \mapsto c, \quad c \in C \tag{2.1}$$

where C is a predefined set of categories (e.g., "dog," "cat," "heart").

For VLMs, the labels can be extended to free text:

$$I \mapsto t, \quad t \in T \tag{2.2}$$

where T denotes the space of free-form text strings, such as captions or question-answer pairs.

Bounding Box Annotations At a more localized granularity, bounding box annotations are extensively used in object detection tasks. Bounding boxes define rectangular regions around objects of interest within an image I, thereby capturing the pixels of the object as well as some additional background pixels. While these annotations provide precise localization, they inherently lack exact boundary detail, which can sometimes limit accuracy in tasks that require high spatial precision. A bounding box B is defined by four coordinates, where (x_{min}, y_{min}) represents the coordinates of the top-left corner and (x_{max}, y_{max}) represents the coordinates of the bottom-right corner. Formally, bounding box annotations map an image to a finite set of labeled rectangles:

$$I \mapsto \{(B_k, c_k)\}_{k=1}^{K}, \quad B_k = \left(x_{\text{min}}^{(k)}, y_{\text{min}}^{(k)}, x_{\text{max}}^{(k)}, y_{\text{max}}^{(k)}\right), \quad c_k \in \mathcal{C} \tag{2.3}$$

Here, $\mathcal{C}$ denotes the finite set of predefined object categories, $K \in \mathbb{N}$ is the number of annotated objects in image I, and all coordinates are assumed to lie within the discrete image domain $\{1, \ldots, W\} \times \{1, \ldots, H\}$.

Pixel-level Annotations For the highest granularity, pixel-level annotations are employed, particularly in semantic segmentation and instance segmentation tasks. Semantic segmentation annotations on the other hand, separate object boundaries at the pixel level but do not differentiate between individual instances; for example, two adjacent dogs would be labeled together without differentiation. Conversely, instance segmentation annotations contain additional information that distinguishes separate instances, labeling each object uniquely (e.g., differentiating "dog 1" from "dog 2"), thus enabling more detailed analyses and validation. Formally:

For semantic segmentation, a segmentation mask S maps each pixel position (i, j) in an image of size $H \times W$ to a class label from a set of categories $\mathcal{C}$:

$$S : \{1, \ldots, H\} \times \{1, \ldots, W\} \mapsto \mathcal{C} \tag{2.4}$$

For instance segmentation, the mask S additionally assigns an instance identifier from the set of natural numbers $\mathbb{N}$:

$$S : \{1, \ldots, H\} \times \{1, \ldots, W\} \mapsto \mathcal{C} \times \mathbb{N} \tag{2.5}$$

An overview illustrating these different annotation types is provided in Figure 2.3.

Overall, annotations must align closely with the specific task requirements. With the ongoing shift toward FMs, tasks have become broader and more general-purpose, increasing the diversity of annotations needed to support comprehensive validation. Nevertheless, within benchmarking contexts, annotations remain critical in precisely assessing and highlighting specific capabilities and limitations of the evaluated models [64].

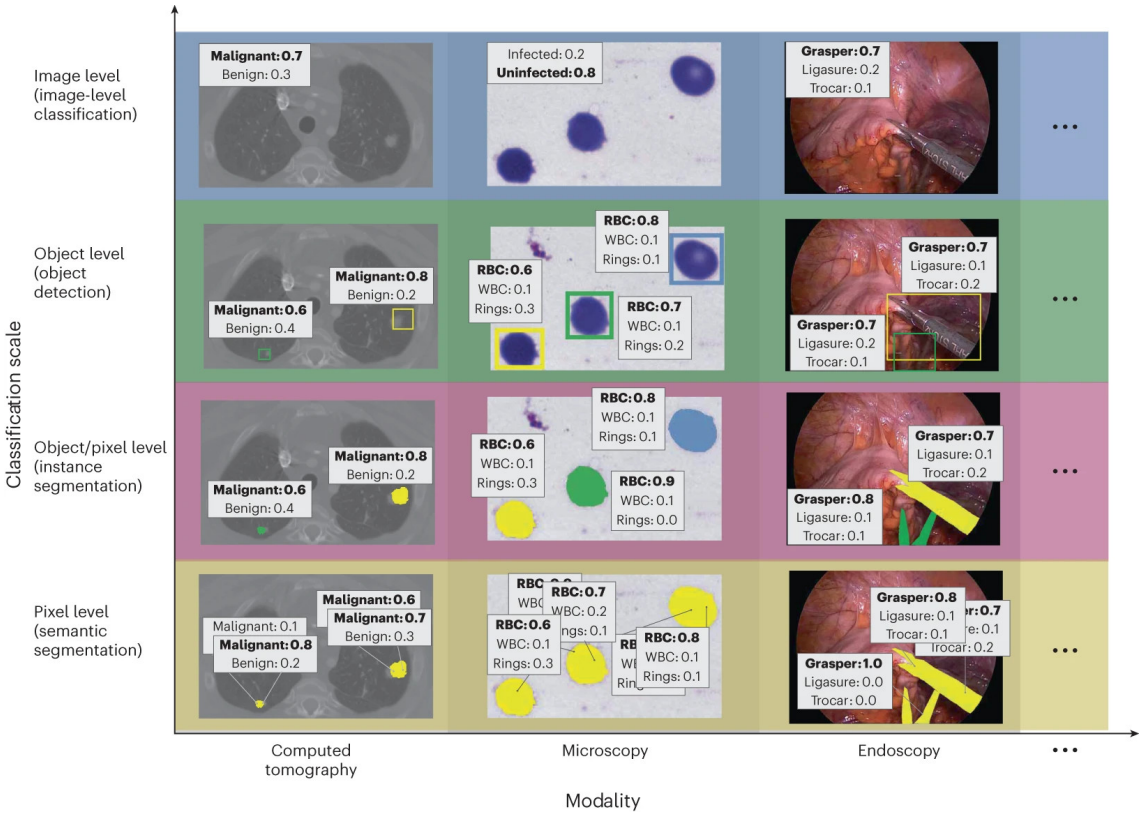


Figure 2.3: **Overview of the different annotation types from image to pixel level.** Whereas classification provides the least information at the image level, instance segmentation provides the most detailed information by combining object- and pixel-level annotations. The figure was originally published in [118].

2.2.2 Validation through Benchmarks

Benchmarks are a foundational component of driving and measuring progress across research disciplines. Broadly speaking, a benchmark is a standardized test or set of criteria used to evaluate and compare the performance of different methods, models, or systems under controlled, repeatable conditions [184]. When researchers develop a new approach, be it a VLM, an algorithm,

or any other solution, they typically rely on benchmarks to measure how their method performs in comparison to existing ones. In the following, we will describe the key components of a benchmark.

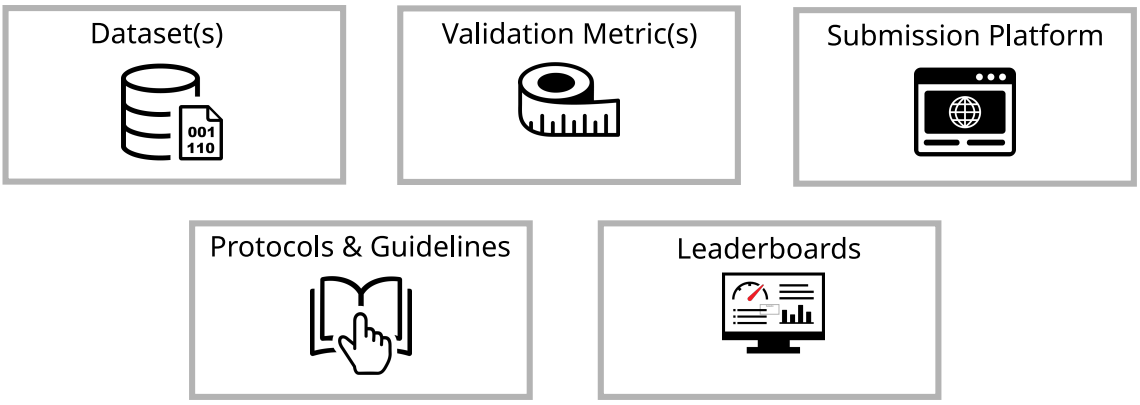


Figure 2.4: **Key components of a benchmark.** Benchmarks typically consist of (i) *datasets*, which provide the data points; (ii) *validation metrics*, which quantify performance; (iii) a *submission platform*, often containerized to upload and run the algorithms; (iv) *protocols & guidelines*, which regulate data usage, resources, and eligibility; and (v) *leaderboards*, which record and transparently compare performance.

2.2.3 Key Benchmark Components

Dataset(s)

Benchmarks revolve around carefully selected datasets containing observations (data points) relevant to the problem the benchmark aims to solve [184]. These datasets can be synthetic, real, or a combination of both, ideally capturing a broad range of conditions (e.g., varied image types, languages, or domains) to reflect real-world scenarios. Each data point is typically annotated or labeled according to a consistent scheme, ensuring fair comparisons across methods.

Validation Metric(s)

Metrics should be selected to align with the benchmark's core objectives, providing standardized measures to quantify performance. Consistent application of these metrics enables researchers to compare and optimize their methods effectively. Employing a combination of metrics can be particularly beneficial, as it balances various validation aspects. For a comprehensive overview of metric-related pitfalls in image analysis, we refer the reader to Reinke et al. [161] and Maier-Hein et al. [118] for metric selection strategies.

Submission Platform

In earlier benchmarks, participants would download test cases, run their algorithms locally, and then manually upload the results. Today, the majority of submissions are typically made via Docker, a containerized environment that packages the code, dependencies, and runtime settings into a single unit. The organizer can be defined as the entity that provides the data, sets the evaluation rules, and oversees the benchmarking process. This approach not only ensures

reproducibility by providing the organizer with the complete code and environment, but also prevents issues such as data alteration or unauthorized external resources. In contrast, merely submitting results would leave organizers without access to the underlying code, making it difficult to verify and reproduce the findings.

Protocols and Guidelines

Benchmarks specify clear rules regarding data usage, submission instructions, external resources, and constraints on time and compute resources for each submission, among others. These guidelines ensure a level playing field by preventing overly specialized solutions or unauthorized private data. The protocols also define eligibility criteria for participants and outline the overall data structure to be used, ensuring that all submissions follow a consistent format [115]. Additionally, benchmarks may offer prize money or awards to recognize outstanding performance and incentivize innovation.

Leaderboards

Leaderboards provide a transparent record of performance, allowing participants to see where they stand relative to others. This not only motivates continuous improvement but also helps the research community identify high-performing approaches worthy of further exploration. In many cases, especially for challenges (see Section 2.2.4 for a definition of the term), benchmarks feature both a public leaderboard offering real-time feedback and a private leaderboard used for final validations. This dual approach ensures that while participants can monitor their progress during the competition and validate their submission code's functionality, the final rankings remain robust and less biased.

2.2.4 Benchmarks vs. Challenges

While a benchmark provides a stable, ongoing validation framework, a challenge typically revolves around a focused, time-bound competition. Challenges have most of the time a defined start and end dates, pushing participants to innovate and achieve the highest possible scores or breakthroughs within a short timeframe, often with prize money or awards to recognize outstanding performance [115]. After the competition concludes, it is common for organizers to keep the submission portal open, allowing researchers to continue submitting new or improved algorithms. Over time, popular challenges can transition into de facto benchmarks when their datasets and validation methods become widely adopted. For example, ImageNet [94] has gathered over 80,000 citations since running its initial challenge. Meanwhile, benchmarks themselves remain relatively fixed and serve as a consistent measure, ensuring that results from different methods are directly comparable over the long term.

2.2.5 Benchmark Design Decisions

Designing a benchmark is a challenge in itself, as every design decision can significantly affect downstream results [115]. Each component of a benchmark, as outlined in this section, must be carefully selected to ensure the benchmark's overall integrity. Nevertheless, benchmarks are subject to several common limitations, two of which are overfitting and the use of unrepresentative metrics [118].

Overfitting occurs when researchers develop architectures or training protocols overly specialized to the benchmark data. Although such solutions may achieve record-breaking scores, they often fail to generalize to new tasks or datasets. This issue is exemplified by the widespread use of ImageNet, which initially served as a pivotal benchmark in computer vision; incremental improvements today on ImageNet [94] do not always lead to broader real-world progress [189].

Similarly, optimizing methods based solely on the benchmark's chosen metrics can lead to misleading conclusions about performance. For example, in medical imaging, the Dice Similarity Coefficient (DSC) is frequently used to evaluate models, yet this overlap-based metric does not fully capture complex clinical needs in practice [161].

2.3 Deep Learning

2.3.1 Data Splits

Datasets are commonly partitioned into three subsets: training, validation, and test sets [128].

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}}, \quad \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{val}} \cap \mathcal{D}_{\text{test}} = \varnothing \tag{2.6}$$

$$|\mathcal{D}_{\text{train}}| : |\mathcal{D}_{\text{val}}| : |\mathcal{D}_{\text{test}}| \approx \alpha : \beta : \gamma, \quad \alpha + \beta + \gamma = 1 \tag{2.7}$$

The training set is used to optimize the model's internal parameters. The validation set serves to tune hyperparameters and monitor performance during training, often guiding early stopping to prevent overfitting. The test set is held out entirely during training and is used solely for final validation. Standard split ratios for $\alpha : \beta : \gamma$ typically range from $70/15/15\%$ to $80/10/10\%$, though exact proportions depend on specific circumstances, such as the dataset size and domain conventions.

In some cases, k-fold cross-validation is used to enhance statistical robustness, especially for smaller datasets [169]. For k-fold cross-validation, the dataset is partitioned into k equal-sized folds. The model is trained and evaluated k times, each time using a different subset as the validation set while the remaining subsets are used as the training data. The selected metrics are computed by averaging over the k validation sets:

$$\text{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{M}(\mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{val}}^{(i)}), \tag{2.8}$$

where $\mathcal{M}$ represents the selected metric computed for the i-th fold, $\mathcal{D}_{\text{train}}^{(i)}$ denotes the training set for fold i, and $\mathcal{D}_{\text{val}}^{(i)}$ denotes the validation set for fold i.

Overall, data splitting and cross-validation protocols aim to approximate the generalization to real-world data, mitigate overfitting, and generate performance estimates that accurately reflect a model's capability on unseen data.

2.3.2 Optimization and Learning

Training a supervised deep neural network involves iterative optimization of model parameters to minimize a loss function. The loss function measures the difference between the predicted outputs and the ground truth annotations. For classification tasks, the *categorical cross-entropy* loss is commonly used [120], and is formally defined as

$$\mathcal{L}_{\mathsf{CE}}(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{2.9}$$

where $y = (y_1, \ldots, y_C)$ denotes the one-hot encoded ground truth vector, $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_C)$ is the predicted class probability vector output by the model, and C is the number of classes.

For segmentation tasks and the biomedical domain in particular [224], the *Dice loss* is often employed, defined as

$$\mathcal{L}_{\mathsf{Dice}}(y, \hat{y}) = 1 - \frac{2\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i}, \tag{2.10}$$

where y_i and $\hat{y}_i$ represent the ground truth and predicted binary label, respectively, for pixel i.

The optimization is typically performed using stochastic gradient descent (SGD) [128] or one of its adaptive variants, such as Adam [90]. The standard approach employs mini-batch training, where gradients are computed over small subsets (mini-batches) of the training data, rather than the entire dataset, and model parameters θ are updated iteratively.

A complete pass through the entire training set is referred to as an epoch, and training is performed over multiple epochs to progressively refine the model parameters. Key hyperparameters include the learning rate η and (mini-) batch size, both of which affect the convergence behavior, optimization stability, and ultimately the model's performance.

2.4 Convolutional Neural Networks

DL is a subset of Machine Learning (ML). ML comprises algorithms that improve automatically through experience [128]. Within the broader field of Artificial Intelligence (AI), Convolutional Neural Networks (CNNs) are to this day a pivotal technology in computer vision. A comprehensive historical overview of neural network developments is presented by Schmidhuber [170]. Below we provide a brief overview of key historical milestones of CNNs and the transition to transformer-based architectures.

2.4.1 Key Convolutional Neural Network Architectures

The first major breakthrough in CNNs is commonly attributed to AlexNet (2012) [94], which won the ImageNet Large Scale Visual Recognition Challenge with a substantial performance gap of ten absolute error points. AlexNet demonstrated the potential of end-to-end learning using convolutional architectures, displacing manually engineered image descriptors that dominated pre-2012 approaches [167].

Following AlexNet, subsequent developments focused on increasing model depth. VGGNet [178], introduced in 2014, extended the architecture to 16 and 19 layers, using uniform 3×3 convolutions

instead of the 7×7 convolutions used in AlexNet. However, increasing depth introduced optimization challenges, such as vanishing gradients. ResNet [65], introduced in 2015, addressed this by incorporating residual connections that enable gradient flow, enabling training of architectures with over 100 layers. DenseNet [74], proposed in 2016, further enhanced the connectivity by feeding feature maps from all previous layers into each subsequent layer, improving the parameter efficiency and feature reuse. In parallel, several architectures focused on computational efficiency and scaling. The Inception family used parallel convolutions to capture multi-scale features. Later, models like MobileNet and EfficientNet [182] optimized for speed and parameter efficiency, with EfficientNet introducing a compound scaling method that balances image resolution, the model depth and model width.

2.4.2 Domain-Specific Adaptation and Emergence of Alternative Architectures

As CNNs matured, domain-specific adaptations and alternative architectures emerged. In biomedical imaging, U-Net [165] was introduced for semantic segmentation tasks. U-Net combines a contracting encoder with an expansive decoder, linked by skip connections that preserve spatial information at various resolutions, making it well-suited for applications with limited annotated data to this day.

More recently, alternative vision architectures have been proposed. Vision Transformer (ViT) [45], introduced in 2020, replaces convolutional operations with an attention mechanic focused architecture that dispenses recurrences and convolutions. ViT divides an input image into fixed-size non-overlapping patches, flattens and linearly projects each patch into one trainable vector, which is then processed through several self attention blocks. This process is analogous to word processing in Natural Language Processing (NLP) with the original Transformer model [45] and the key differences for both architectures are displayed in Figure 2.5. This enables global context modeling and long-range dependency capture, which can outperform CNNs on large-scale benchmarks when sufficient data is available. A fair comparison between model architectures requires careful control of numerous design choices, as reported rankings are often misleading. Factors such as available Video Random Access Memory (VRAM), the volume and quality of training data, and differences in baseline implementations can substantially influence performance assessments and may confound the interpretation of comparative results [81].

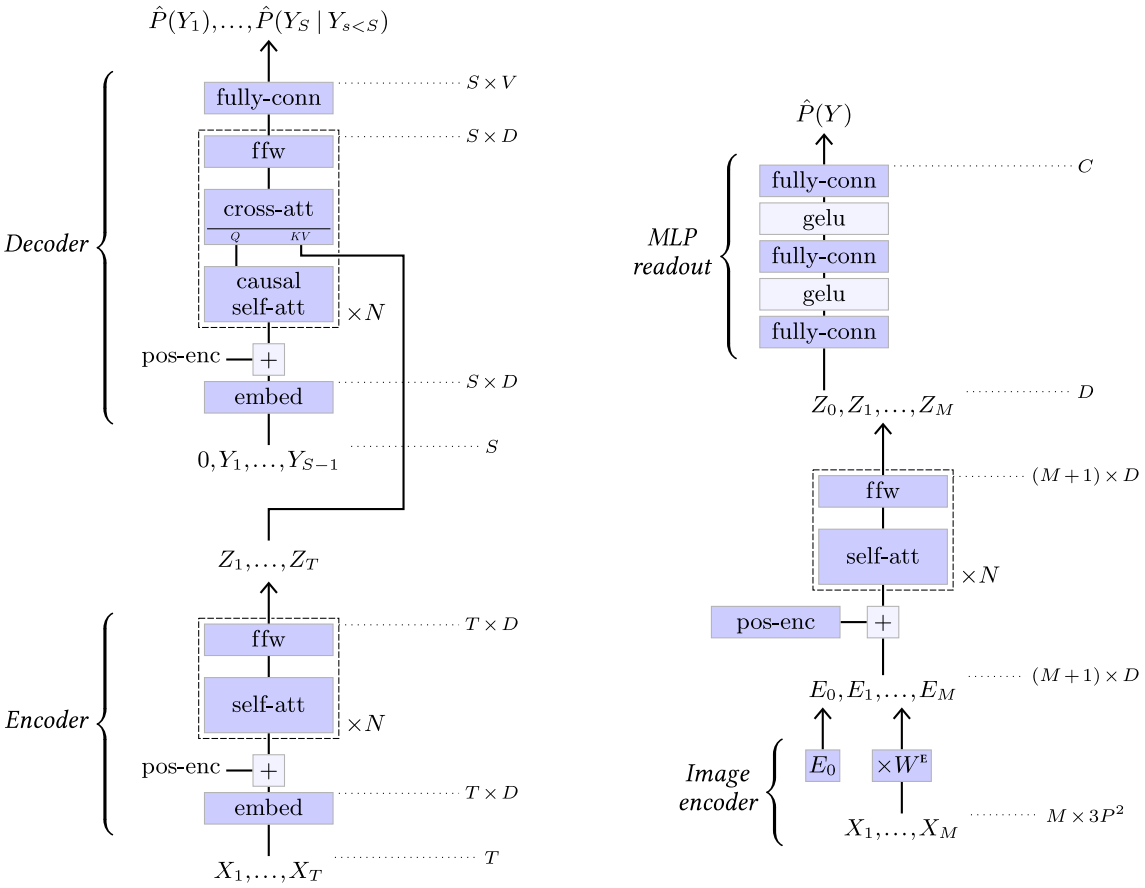


Figure 2.5: **The Transformer architecture and its adaptation to the vision domain, the Vision Transformer.** The original Transformer architecture (left) is designed for sequential data, such as text, and consists of an encoder-decoder structure. The encoder maps input tokens into contextualized representations, while the decoder generates the output sequence using self-attention, cross-attention, and feed-forward layers. The Vision Transformer adapts this concept for image analysis by separating the image into fixed-size non-overlapping patches, flattening and linearly projecting each patch into one trainable vector, which is then processed through several self attention blocks. Illustrations adapted from [49] (Original license: CC BY-NC-SA 4.0).

2.5 Vision-Language Models

Trend toward Multimodality

Historically, ML models addressed a single modality [170]. For instance, as outlined in the previous discussion of CNNs, vision-specific networks achieved impressive performance in image classification and object detection where only pixel-based input was analyzed [170]. Likewise, Large Language Model (LLM) operated exclusively on text. Yet, recent years have witnessed a remarkable shift toward multimodal architectures that integrate and jointly process information from multiple modalities [6].

This trend toward multimodality in the form of VLMs is motivated by the way humans naturally perceive and reason about the world, where they merge multimodal inputs such as sight, sound, and language [220]. In ML, recent progress has largely focused on the integration of two core modalities: vision and language [6]. It can be expected that over time models are extended to incorporate additional modalities, such as audio or depth, supporting a more holistic and context-aware understanding.

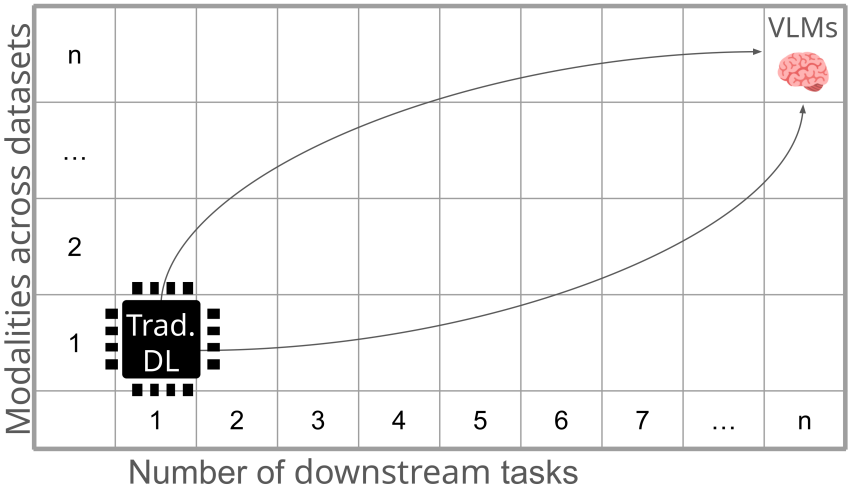


Figure 2.6: **VLMs emerged from single task and single modality DL Models.** While each existing model can be placed in the grid, the trend pushes towards models covering multiple modalities and tasks. As datasets scale, so do annotation errors. Annotation errors persist in narrow and intensify within multimodal datasets. Expanding the range of downstream tasks and modalities compounds the risk of annotation errors, underscoring the value of rigorous data curation.

From Specialized Models to Foundation Models

In parallel to the rise of multimodality, the field has progressed from task-specific models to FMs that provide architectures capable of solving a wide range of tasks with a single model. Traditional DL Models were typically trained to perform one narrowly defined task, such as image classification, object detection, or segmentation for a defined range of categories. These models initially relied on supervised learning with explicit class labels, as seen in benchmark datasets like CIFAR [135], and later ImageNet [167], each with fixed and predefined categories. The definition of the term *Foundation Model* remains subject to substantial interpretative variability within the scholarly community. Bommasani et al. [14] underscore this lack of terminological precision, characterizing *“foundation model”* as an *"informal designation encompassing a broad and evolving family of architectures. The scope and criteria for inclusion are inherently dynamic, reflecting ongoing advancements and paradigm shifts in the field."* Given the inherent vagueness of the term, and recognizing that the VLMs examined in this thesis are applied across a diverse range of tasks, we adopt the convention of classifying each VLM under consideration as a FM for the purposes of this work.

Core Architectural Paradigms

Recent work on VLMs centers on three paradigms.

1. **Encoder-only models**, exemplified by Contrastive Language-Image Pretraining (CLIP) [150], map both images and text with separate encoders into a shared embedding space, rendering them particularly effective for retrieval and similarity-based tasks.
2. **Encoder-decoder models**, such as BLIP-2 [105], combine a visual encoder with a language decoder, enabling tasks that require grounded text generation, including image captioning and VQA. Hybrid approaches such as Flamingo extend this paradigm by augmenting pretrained decoder-only LLMs with cross-attention layers to visual features[6].
3. **Unified transformers**, for instance VL-T5 [27], merge encoding and decoding within a single transformer stack, facilitating multitask training with shared parameters.

Contrastive Language-Image Pre-training (CLIP)

As highlighted, an influential advance enabling foundation models for vision-language tasks was CLIP [150]. Instead of using fixed class labels, CLIP exploits widely available, naturally occurring internet data comprising image-caption pairs. This vastly expands the concepts and diversity present in the training data, allowing models to represent the visual world far beyond fixed-class datasets. The corresponding softmax contrastive loss is

$$-\frac{1}{2N}\sum_{i=1}^{N}\Big(\log\frac{e^{\mathrm{sim}(\mathbf{I}_i,\mathbf{T}_i)/\tau}}{\sum_{j=1}^{N}e^{\mathrm{sim}(\mathbf{I}_i,\mathbf{T}_j)/\tau}}+\log\frac{e^{\mathrm{sim}(\mathbf{I}_i,\mathbf{T}_i)/\tau}}{\sum_{j=1}^{N}e^{\mathrm{sim}(\mathbf{I}_j,\mathbf{T}_i)/\tau}}\Big), \tag{2.11}$$

where $\mathbf{I}_i$ and $\mathbf{T}_i$ denote image and text embeddings for the i-th pair, sim is cosine similarity, τ is a temperature parameter, and N is the batch size.

Building upon CLIP, Sigmoid-based Contrastive Loss (SigLIP) [219] replaced and simplified the matrix-based softmax with a pointwise sigmoid formulation, improving computational efficiency and representation quality:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\Big(\frac{1}{1+\exp\big(z_{ij}(\langle\mathbf{l}_i,\mathbf{t}_j\rangle/\tau+b)\big)}\Big), \tag{2.12}$$

with $\mathbf{l}_i, \mathbf{t}_j$ the embeddings, $\langle\cdot,\cdot\rangle$ the inner product, $z_{ij}\in\{-1,1\}$ the pair label, τ the temperature, and b the bias term.

However, a shortcoming of contrastive learning lies in its training objective: it only requires distinguishing matching image-text pairs from mismatched ones. As a result, the model learns to capture just enough information to make this distinction and consequently often ignores fine-grained details that are not strictly necessary to resolve the contrastive task. For instance, consider an image of a human standing next to a dog. A contrastively trained model might only learn to associate the image with generic phrases like “a person and a dog,” without capturing specific attributes (e.g., the color of the dog), spatial relations (e.g., “a human on the left and a dog on the right”), or interactions (e.g., “petting the dog”). These details are not incentivized by the contrastive loss, and thus are typically absent from the learned representation. The issue is displayed in Figure 2.7.

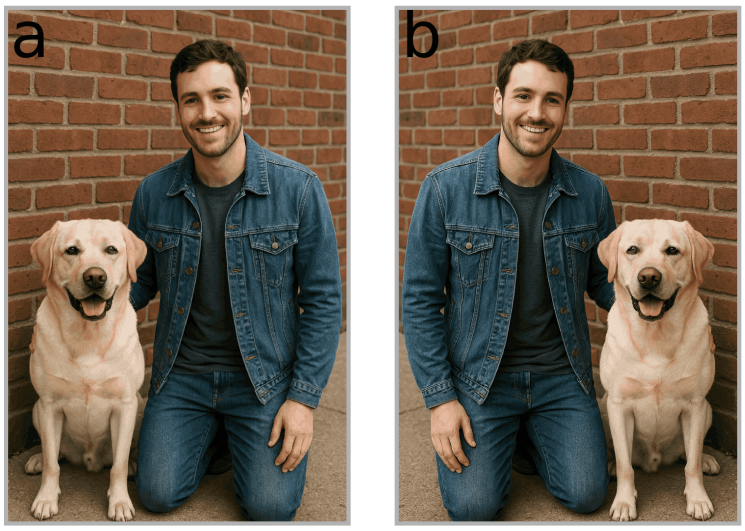


Figure 2.7: **CLIP does not capture spatial relationships between objects**. In both images, a human and a dog appear together; in (a), the dog is to the left of the human, while in (b), the dog is to the right. A contrastively trained CLIP Model would treat these images as equivalent, encoding only the presence of "a person and a dog" and ignoring spatial arrangement or object interaction. Base image generated with generative AI.

Beyond Contrastive Learning

To address the limitations of contrastive learning, recent research emphasizes the importance of explicitly encoding detailed textual descriptions via image captioning [187]. Caption-based approaches overcome the sparse relational information of contrastive training by directly encoding attributes, object relationships, and contextual information that are present in the captions, resulting in richer visual-textual representations. Consequently, modern VLMs are evolving from models trained solely with contrastive objectives to ones integrating more structured textual representations.

Data and Annotation Challenges

The training and benchmarking of VLMs face unique challenges related to multimodal data collection and annotation in two dimensions, as depicted in Figure 2.6. First, multimodal datasets require pairing data across modalities, which complicate the sourcing, scaling, and annotation of the obtained data [198]. Utilizing large-scale, unstructured internet data introduces risks of inherent biases, such as skewed representation or stereotypical associations, potentially impacting model fairness and generalization [5]. Addressing these challenges requires extensive dataset curation and bias-mitigation strategies to ensure robust, fair, and generalizable multimodal models [171]. Second, the growing diversity of downstream tasks a model is expected to handle complicates performance assessment [198]. This necessitates careful balancing of task-specific performance and underscores the importance of performing dedicated evaluations tailored explicitly to individual tasks [161].

3 Related Work

In this chapter, we describe the related work of this thesis. We start with the topic of data annotation in Section 3.1, where we establish a taxonomy of annotation methods ordered by decreasing human involvement. Subsequently we introduce QA and a taxonomy of QA methods ordered by decreasing human involvement in Section 3.2. Lastly we shift the focus in Section 3.3 to VLM benchmarks, where we highlight the historical development of benchmarks along major algorithmic achievements and end with recent topics related to QA in benchmarks.

3.1 Data Annotation Methods and Challenges

3.1.1 Data Annotation Introduction

The process of data annotation can be defined as the process of systematically assigning labels or metadata to raw data, which can encompass diverse sensor outputs such as images, text, or video [155]. It constitutes a foundational pillar for test data in general, and additionally for training data for supervised and semi-supervised ML. In contemporary research contexts, especially those involving VLM, the annotation process has grown both in methodological sophistication and in complexity, introducing a host of technical, logistical, and epistemological challenges [198], as highlighted in the last chapter. In this section, we critically review prevailing annotation methods, display inherent difficulties, and situate these within the broader discourse on model validation and benchmarking. Each subsection concludes with an explicit link to the Research Question (RQ) it motivates or addresses.

3.1.2 Development of Annotation Markets

The annotation market has not remained static but has developed along the steadily growing requirements in AI [197]. We can distinguish four characteristic phases: early manual annotation conducted in academic laboratories, the rise of large-scale crowdsourcing platforms, the subsequent professionalization of annotation companies, and the establishment of expert-driven annotation in specialized domains. Each of these phases was necessary to progress to the next phase.

Manual Annotation in Academic Research Laboratories

In the early phases of AI research, annotation was carried out almost exclusively by researchers, their students, or project collaborators within academic laboratories [100]. The datasets and models developed[1] were typically small, often comprising only a few thousand samples, which made fully manual annotation feasible despite its labor-intensive nature. The tasks of this period primarily focused on classification, which requires the least amount of annotation effort per sample compared to more complex tasks. One of the most prominent datasets of that era, MNIST [100], consisted of 60,000 28×28 images across 10 classes. Scaling up the number of classes and resolutions initially reduced the number of samples available per class, exemplified by Caltech-101 [104], which included 101 classes and a total of 9,146 images, with between 40 and 800 images per class and a resolution of 300×200 pixels. The limited scale of these annotation efforts was sufficient for the comparatively simple models at that time, and many of the resulting datasets went on to serve as foundational benchmarks for the community.

Crowdsourcing Platforms

With the expansion of ML in the 2000s and the growing demand for larger datasets, annotation could no longer be completed within a reasonable time frame by academic laboratories alone. As a response, annotation tasks began shifting to general crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk). Introduced in 2005 and named after the eighteenth-century chess-playing automaton "Mechanical Turk" [180], MTurk provided access to a large pool of laymen workers. See Figure 3.1 for an illustration of the Mechanical Turk. On these platforms, users could upload any type of task, referred to as Human Intelligence Tasks (HITs), ranging from writing a poem or searching for an address to annotating data points for ML [44]. This business model enabled rapid scalability at relatively low cost by providing access to a broad and worldwide user base.

Driven by the growing needs of the ML community, the first marketplace dedicated specifically to data annotation began operating in 2007 [35]. Over the past decade, MTurk has remained the predominant choice in academia [32]. This demand is exemplified by ImageNet [167], arguably the most influential academic dataset annotated via MTurk. It consists of more than 20,000 categories and 14 million images, with annotations collected over roughly 1.5 years from more than 49,000 crowdworkers across 150+ countries. However, MTurk's incentive structure has proven problematic in recent years [157]. Low pay [111] and uneven power dynamics [41] often led workers to prioritize speed over accuracy on their annotation tasks in order to maximize their earnings. These trends have been linked to a decline in annotation quality on MTurk in recent years [89].

[1] At least by today's standards in 2025.

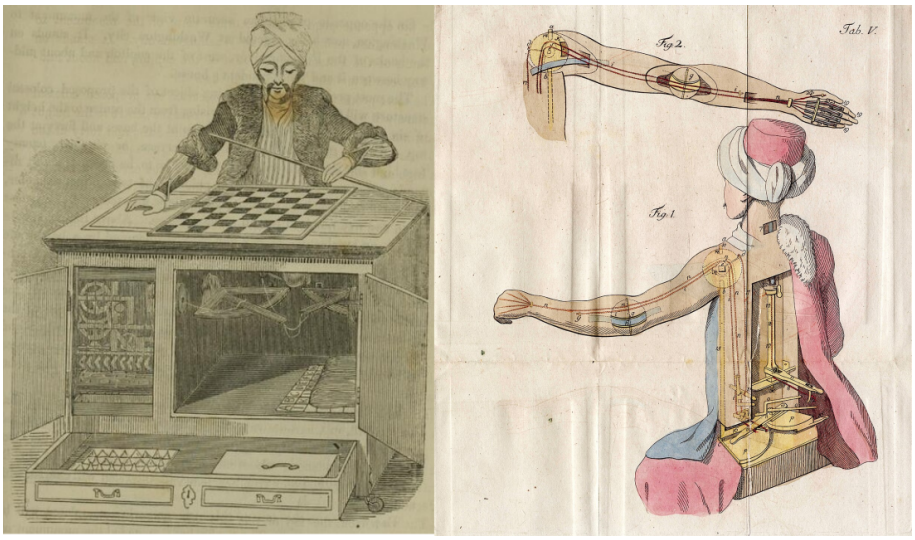


Figure 3.1: **Amazon Mechanical Turk is named after the chess robot created by Wolfgang von Kempelen.** While it was presented to the observer as an automated machine, it secretly concealed a human operator inside. The machine was constructed by Wolfgang von Kempelen. Amazon's MTurk platform later borrowed its name from this illusion of automation. Images taken from [50, 229] (Both images are in the public domain).

Professional Annotation Companies

With the ever growing need for annotations, the last decade witnessed the emergence of professional annotation companies, such as Appen, Scale AI, and others [11, 117]. Unlike crowdsourcing platforms, these companies specialize exclusively in data annotation and provide institutional scaffolding for annotation work, including structured training programs and internal QA procedures. On crowdsourcing platforms, by contrast, the responsibility for task design, training, and tooling lies entirely with the requester.

From an employment perspective, professional annotation companies predominantly rely on full-time, on-site annotators [155]. In the beginning, the employed people were mostly laymen, which resulted in relative low pay for the annotators. Yet, the distinction between crowdsourcing platforms and professional annotation companies has begun to blur in recent times, as larger providers increasingly accommodate diverse employment models [157]. For certain application domains, such as projects commissioned by the United States Government, full-time employment and dedicated security clearance are mandatory for annotators [205].

To demonstrate their expertise, professional annotation companies have increasingly begun to release proprietary datasets. These take the form either of improved versions of existing resources or entirely new datasets. An example of the former is the SAMA-COCO dataset, a relabeling of COCO-2017 [109] that added approximately 25% additional object instances to the existing images [227]. An example of the latter is the nuScenes dataset, a large-scale autonomous driving dataset with multimodal sensor data and 3D object annotations [18].

Expert-driven Annotation in Specialized Domains

With the rapid rise of generalist and reasoning models, new benchmark high scores appeared almost every week [64]. Performance often reached the internal error rate of the datasets, which led many to consider them solved and made it increasingly difficult to design new benchmarks [135, 19]. This situation led to the introduction of highly specialized benchmarks that target narrow areas of domain knowledge. A prominent example is Humanity's Last Exam [146], built to test VLMs at the edge of human academic expertise. It consists of 2,500 expert-authored questions covering mathematics, humanities, and natural sciences, with answers that are unambiguous, verifiable, and designed to resist simple retrieval. The data was collected through a worldwide crowdsourcing campaign combined with a prize pool to incentivize participation of domain experts.

Annotation companies faced a similar shift in project requirements. Depending on the project and the data, customers increasingly demanded not lay annotators but domain experts. These experts were more difficult to recruit, and the power dynamic changed, where annotators with specialist knowledge could command far higher wages, in some cases up to 85 USD per hour [3], and better working conditions as opposed to laymen annotators. This development has given rise to a range of startups focused on sourcing specialized data for leading laboratories. Some recruitment companies, such as Mercor AI [2] or Turing [190] have likewise begun to pivot or extending their business toward providing domain experts directly to major AI labs.

Given the development trajectory, these new annotation requirements are unlikely to be temporary. Comparable trends can already be seen in robotics, where an increasing number of companies are working on recording highly specialized training data [158].

3.1.3 Taxonomy of Annotation Methods

The annotation of data for ML, particularly in the context of computer vision and multimodal systems, can be differentiated according to the extent of manual human interaction within the process. We differentiate annotation methodologies as placed along a continuum of decreasing human intervention, ranging from wholly manual procedures through hybrid (model-assisted) regimes to fully automated systems. This differentiation is not merely of theoretical interest, but reflects practical implications for annotation quality, scalability, and reproducibility. Given the high monetary and time cost resulting from human interaction, researchers and practitioners try to minimize the human involvement as much as possible. The annotation methods overview is displayed in Figure 3.2.

Manual Annotation

Manual annotation remains the gold standard for generating high-fidelity, contextually accurate labels, particularly in domains where expert judgment is indispensable (e.g., medical imaging and surgical data [117]) and for generating test data in general. Typical workflows include instruction development, annotator training, and iterative feedback, see Section 2.1.2. Several sub-variants can be distinguished:

- **Expert Annotation:** In scenarios involving small or highly specialized datasets, annotation may be conducted by domain experts whose decisions serve as the reference standard. This is in particular the case in biomedical image annotation [83].

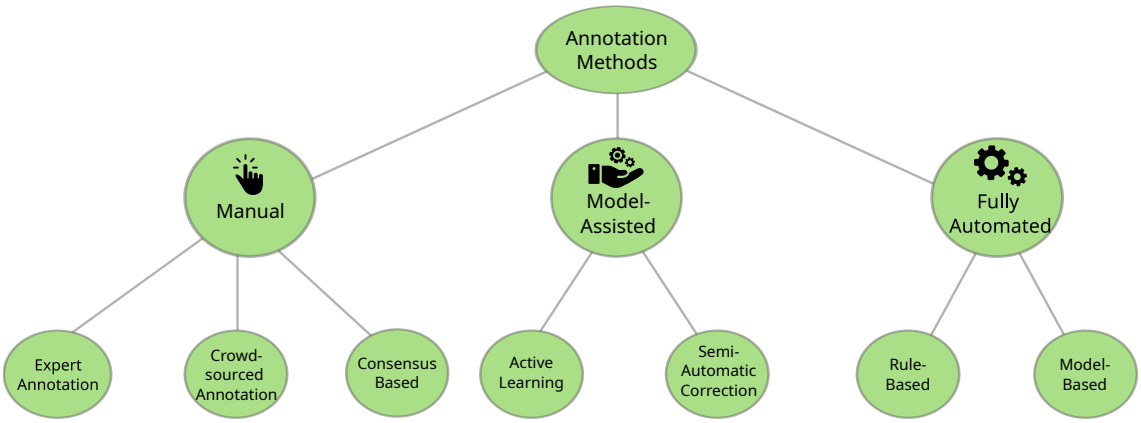


Figure 3.2: **Taxonomy of annotation methods ordered by decreasing human involvement (left to right).** The hierarchy illustrates three principal families from Manual over Model-Assisted to Fully Automated Annotation together with their common sub-variants. The left-to-right order indicates a progressive reduction in human involvement.

- **Crowdsourced Annotation:** To achieve scalability, annotation tasks are often distributed to large pools of non-expert annotators via online platforms such as Amazon Mechanical Turk or professional annotation companies. In contrast to Amazon Mechanical Turk, an annotation company specializes in the production of labeled data, their annotation personnel are predominantly employed on a full-time basis and operate within conventional office environments. While crowdsourcing increases throughput, it may also introduce greater label variability and noise [141].
- **Consensus-Based Annotation:** To mitigate individual annotator bias and enhance label quality, multiple annotations are aggregated through majority voting, probabilistic modeling (e.g., the Dawid-Skene algorithm [34]), or weighted aggregation schemes, producing a consensus label that is presumed to approximate the reference annotation. It should be noted that some part of community considers this QA rather than annotation.

Model-Assisted Annotation

Progressing along the continuum toward reduced human input, model-assisted annotation (sometimes termed semi-automatic annotation in older publications [13]) integrates ML models into the labeling workflow to enhance efficiency. Here, pre-trained or incrementally updated models propose candidate labels or highlight uncertain cases, with human annotators retained in the loop to verify, correct, or refine these suggestions. A major drawback of these methods is automation bias, in which users tend to accept a model's suggestions regardless of their correctness [179].

Key methodological variants within this category include:

- **Active Learning:** Here, models iteratively select data points for annotation based on estimated uncertainty, expected information gain, or representativeness, thereby prioritizing cases where human input is most informative [73].
- **Semi-Automatic Correction:** Particularly relevant in pixel-level or instance segmentation

tasks, preliminary model outputs (e.g., mask proposals from an nnU-Net [80] in the biomedical domain) are presented to human annotators for refinement. This hybrid approach leverages machine efficiency while preserving human oversight for challenging or ambiguous cases.

Model-assisted annotation offers a pragmatic compromise between scalability and quality, but its effectiveness is contingent upon the initial model's performance and the availability of (expert) annotators for error correction.

Fully Automated Annotation

At the opposite end of the spectrum lies fully automated annotation, wherein automated computational systems perform the labeling process with minimal or no direct human oversight.

- **Rule-Based Labeling:** For relatively simple or well-structured tasks, heuristics and handcrafted rules can be used to assign labels automatically. While efficient, such approaches are inherently limited by their rigidity, lack of adaptability and require deep domain expertise.
- **Model-Based Labeling:** The rise of VLMs, such as CLIP [150], has enabled automated pseudo-labeling of vast image corpora by leveraging multimodal pretraining on web-scale image-text pairs. These models can generate labels or descriptions without task-specific supervision, thus accelerating dataset construction. Another option is to fine-tune a general model on a subset of the collected data with high quality annotations.

While fully automated methods provide unmatched scalability and can address annotation bottlenecks in rapidly growing domains, they are also susceptible to propagating model biases [5], introducing systematic errors [135], and lacking the nuanced judgment characteristic of expert human annotators [174].

Link to this thesis (RQ1)

The rising demand for annotated data and the developments of annotation methods highlight the persistent challenges in obtaining high-quality and consistent annotations. This motivates our investigations of labeling instructions (**RQ1**), where we quantify their importance from the annotator perspective and how labeling instruction design impacts annotation quality.

3.2 Quality Assurance Methods

3.2.1 Quality Assurance Introduction

The main goal of QA is to evaluate and where necessary improve the reliability, accuracy, and usability of generated annotations. Across application domains, a wide range of QA strategies have been developed, differing in cost, scalability, and their ability to detect specific error types [155]. In the context of crowdsourcing platforms, Daniel et al. [33] provide a structured overview, categorizing QA methods by their position in the annotation pipeline (pre-, during-, or post-labeling). Their framework covers a broad taskrange but excludes annotation companies. Lu et al. [113] and Zheng et al. [225] extend the discussion by evaluating post-hoc QA measures performed by experienced personnel from the requester side. However, their analyses are limited in scale and do not investigate QA mechanisms conducted internally by annotation companies.

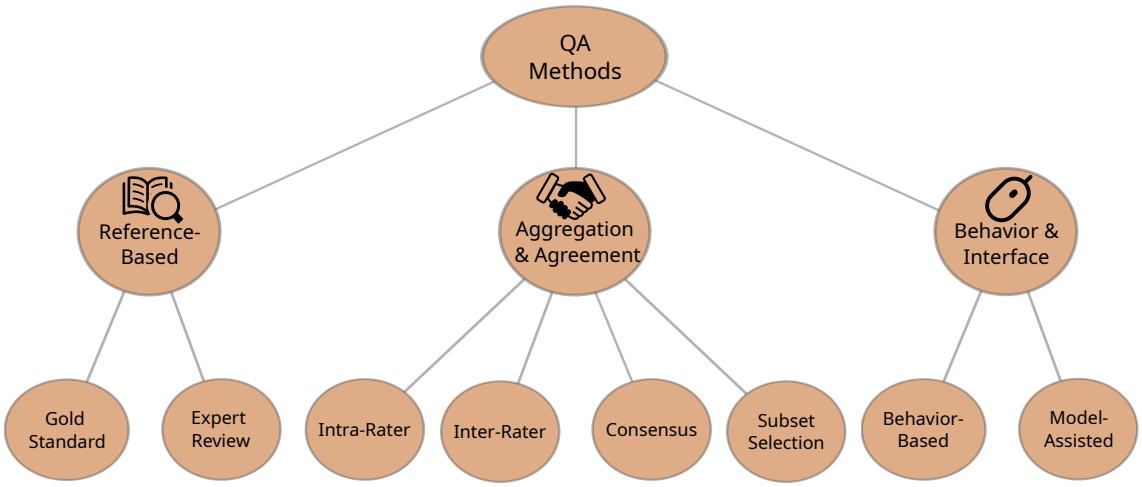


Figure 3.3: **Taxonomy of Quality Assurance methods ordered by decreasing human involvement (left to right).** The hierarchy distinguishes three principal families from *Reference-Based QA*, which relies on direct comparison to trustworthy labels or expert review, over *Aggregation and Agreement*, which derives a quality signal from multiple annotators, to *Behavioral and Interface-Based QA*, which leverages annotator interaction patterns, model-assisted signals, or task design improvements. The left-to-right order indicates a progressive reduction in direct human involvement in the quality assurance process.

3.2.2 Taxonomy of Quality Assurance Methods

A common framework classifies QA methods into three methodological categories [113]:

1. Comparison to a reference, such as gold-standard labels [99]
2. Aggregation and agreement of multiple independent annotations [78]
3. Assessment of annotator behavior or interface interaction [67, 77].

These categories are not mutually exclusive and are often combined in production workflows depending on the annotation project. In the following we describe the major QA methods:

Inter-Rater Agreement Agreement between multiple annotators labeling the same data point is frequently used to assess annotation reliability and less often for instruction clarity. The naive approach calculates the proportion of agreement:

$$p_o = \frac{\text{number of agreements}}{\text{total number of annotations}} \tag{3.1}$$

To account for chance alignment, Cohen's Kappa [118] is widely used for scenarios with two annotators:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.2}$$

where κ is the inter-rater agreement coefficient for two annotators, p_o is the observed proportion of agreement, and p_e is the expected agreement by chance based on the marginal distributions of the ratings.

In contrast to Cohen's Kappa, Fleiss's Kappa [129] can account for scenarios with three or more annotators:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{3.3}$$

where κ is the agreement among n annotators corrected for chance, $\bar{P}$ is the mean observed agreement across all data points, and $\bar{P}_e$ is the mean expected agreement by chance. Note that lowercase p denotes pairwise proportions of agreement (Cohen's Kappa), while uppercase $\bar{P}$ refers to averaged proportions across multiple annotators (Fleiss' Kappa).

$$\bar{P} = \frac{1}{Nn(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{c} n_{ij}(n_{ij} - 1) \tag{3.4}$$

where N is the number of data points, n is the number of annotators per data point, c is the number of categories, and n_{ij} is the number of annotators who assigned data point i to category j.

$$\bar{P}_e = \sum_{j=1}^{c} p_j^2, \quad p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{3.5}$$

where p_j is the overall proportion of ratings assigned to category j.

However, high agreement does not always imply correctness, particularly for inherently subjective or ambiguous tasks [83]. Conversely, disagreement may indicate areas of underspecification in the instructions or multiple plausible interpretations [211].

Intra-Rater Agreement Intra-rater agreement examines whether a given annotator labels the same sample consistently over time. It can reveal temporal instability due to fatigue, insufficient instruction clarity, or cognitive overload. Although relatively underreported, it provides a complementary perspective to cross-annotator agreement and is relevant in both crowdsourced and professional pipelines.

Gold Standard Samples Gold-standard QA compares annotations to expert-generated reference labels [99]. These references are often used for training, calibration, or error auditing. However, they rely on the assumption that expert labels are unambiguous and representative. In tasks with inter-expert variability or low annotation consensus, such gold standards may underrepresent acceptable variance and risk over-penalizing minority interpretations [83].

Expert Annotator Review Secondary review by domain experts serves as a fallback QA mechanism, particularly for difficult or high-risk tasks. This review may target edge cases, ambiguous regions and cases, or final benchmark datasets. While effective, the cost and time required generally preclude its use at scale. In biomedical and legal domains, such review remains common for test set curation or training validation [83].

Consensus Labeling Label fusion through aggregation is a standard procedure in crowdsourced setups. Simple majority voting is defined as:

$$y^* = \arg\max_{c \in \mathcal{C}} \sum_{r=1}^{R} \mathbf{1}\{y_r = c\} \tag{3.6}$$

Since annotators display different reliability, a weighting factor for each annotator can be introduced:

$$y^* = \arg\max_{c \in \mathcal{C}} \sum_{r=1}^{R} w_r \, \mathbf{1}\{y_r = c\} \tag{3.7}$$

where y^* is the consensus label assigned to an data point, $\mathcal{C}$ is the set of possible categories, R is the number of annotators, y_r is the label provided by annotator r, $\mathbf{1}\{y_r = c\}$ is an indicator function equal to 1 if annotator r chose category c, and w_r is the weight assigned to annotator r reflecting their estimated reliability.

The Dawid-Skene model [34] extends this principle by replacing the scalar weight with an annotator-specific confusion matrix, estimating the probability that annotator r reports label y_r given that the correct class is c:

$$y^* = \arg\max_{c \in \mathcal{C}} \prod_{r=1}^{R} \pi_r(y_r \mid c) \tag{3.8}$$

where $\pi_r(y_r \mid c)$ is the entry of annotator r's confusion matrix, representing their conditional reliability for category c.

These methods reduce the influence of outlier annotators and smooth noisy signals, but cannot fully resolve systematic disagreements or subjective ambiguity, which is particularly present in biomedical domain [110].

Statistical Subset Selection To reduce QA costs, annotation reviews are often restricted to a representative or randomly sampled subset of data [107]:

$$\mathcal{S} \subset \{1, \ldots, N\}, \quad |\mathcal{S}| = \rho N \tag{3.9}$$

where $\mathcal{S}$ is the randomly sampled subset of data points selected for review, N is the total number of annotated data points, and $\rho \in (0, 1]$ is the sampling rate.

This allows for the estimation of global error rates or identification of recurrent mistakes. While scalable, subset reviews may miss low-prevalence errors unless stratified carefully or augmented with error-focused heuristics.

Model-Based QA Discrepancies between model predictions and human annotations can be used to flag samples for additional review [225]. This approach is especially valuable when models are well-calibrated and can detect annotation outliers, inconsistencies, or low-confidence cases. It is increasingly used in hybrid pipelines where automated suggestions precede or follow human annotation.

Behavior-Based Auditing Behavioral telemetry data, such as mouse movement, interaction speed, or correction frequency, has been proposed as a proxy for annotation reliability [62, 67]. These signals can be analyzed retrospectively to rank annotators, detect rushed submissions, or identify samples with elevated annotation uncertainty. They are non-invasive and can scale to large datasets with minimal overhead.

Training and Interface Design Preventative QA approaches aim to minimize annotation errors at the source by enhancing annotator comprehension and task execution. Targeted training interventions [46], prototypical feedback cycles [53], and adaptive interface designs [15, 20] have been shown to improve annotation consistency and reduce instruction misinterpretation. While they are not considered QA in the narrow sense, these strategies substantially influence the final annotation quality.

In summary, QA encompasses a broad methodological spectrum, ranging from instruction-level interventions to post-hoc expert review, where most of the literature addresses classification tasks and crowdsourced labeling. QA for complex data types (e.g., dense segmentations), domain-specific pipelines, and internal industrial annotation workflows is relatively underrepresented. Future research should address these gaps by systematically evaluating how QA strategies interact with task ambiguity, annotator expertise, and instruction design under practical constraints.

Link to this thesis (RQ1)

The broad diversity of general QA approaches and yet limited research on internal QA illustrates the shortcomings of current practice. This motivates our investigations of QA (**RQ1**), where we focus on internal QA from annotation companies and specific image metadata.

3.3 Vision-Language Model Benchmarks

3.3.1 Interplay Between Algorithmic Advances and Benchmark Improvements

Historically, major algorithmic achievements go hand in hand with substantial benchmark gains, for example, Bidirectional Encoder Representations from Transformer (BERT) [42] popularized the General Language Understanding Evaluation (GLUE) Benchmark, or Chinchilla [72] strongly promoted the usage of Measuring Massive Multitask Language Understanding Benchmark (MMLU) [70]. Furthermore, the initial and most influential efforts to combine vision and language in benchmarks and foundational models originated within the NLP community [216, 217, 138].

BERT is a representative example with over 100,000 citations since 2018 for the NLP era with a focus on benchmarks such as GLUE. BERT encodes text as a sequence of vectors through self-supervised learning, relying on an encoder-only transformer architecture [42]. Random tokens in the input are masked, and the model predicts the missing tokens based on surrounding context. In contrast to masked self-attention used in decoders, modeling the bidirectional context (both left and right) enables stronger contextual understanding. Its main strength is producing contextual word embeddings for downstream NLP tasks. As an encoder-only model that works by “understanding” context, it is not designed to generate long passages of text in the way GPT-style models do.

In the following, we will exemplarily describe two algorithmic achievements from the NLP research community that shape today’s research landscape and the related benchmarks.

Introduced in 2022, Chain-of-thought prompting [202] claimed a breakthrough on GSM8K [30], which comprises 8,500 high-quality math problems at the grade school level. Chain-of-thought prompting is a technique used in prompting LLMs that encourages them to generate intermediate reasoning steps before generating the final output. Instead of directly producing a final response, the model is guided to 'think aloud' and break down the problem into smaller, sequential steps. This process can lead to more accurate and transparent answers, particularly for complex tasks that require multi-step reasoning or involve intricate logic [202].

HumanEval was introduced and popularized by Codex [23] in 2021 as a benchmark for code generation. Codex [23] is a generative pre-trained transformer language model fine-tuned on over 150 GB of publicly available Python code from GitHub. HumanEval comprises 164 handcrafted programming problems, each evaluated using unit tests to verify correctness. Notably, the authors have observed that fine-tuning GPT-3 does not necessarily yield superior performance compared to training from scratch, and that the BLEU [148] score has significant limitations when assessing the correctness of programming outputs. This underscores a growing recognition of metric limitations, an issue similarly highlighted by Maier-Hein et al. [118] for the biomedical imaging domain.

3.3.2 Development of Benchmark Sizes

The fundamental aim of each data point in a benchmark is to deliver a meaningful signal. As models rapidly surpass existing benchmarks, there is a continuous need for new datasets that can capture increasing complexity [64]. The idea is that increasing the number of data points helps capture more diverse signals and complexities. For example, Multilingual Measuring Massive Multitask Language Understanding Benchmark (MMMLU) now includes over 16,000 questions across 57 fields [70], thereby significantly enhancing the benchmark's capacity to challenge models over its predecessor.

Building on this expansion in size, subsequent efforts have focused on reducing redundancy by identifying representative subsets within these larger benchmarks. We refer to this method as pruning, as less relevant data points are removed to increase the signal of the remaining data points. Recent work on benchmark pruning, such as the development of Mini MMMLU [147], demonstrates that a carefully selected subset, sometimes as small as 100 samples, can yield performance estimates comparable to those obtained from the full dataset. This approach not only accelerates the validation process but also minimizes the environmental impact of the benchmark.

A third trend involves creating new benchmarks with highly curated samples. Initiatives like Humanity's Last Exam [146] exemplify a "less is more" philosophy by employing crowdsourced questions from domain experts that undergo expert review and are intentionally adversarial, ensuring that most current models struggle. One drawback of this approach is that new models may inherently benefit from not being involved in the benchmark's construction phase and thus report a higher performance.

In summary, as benchmarks are quickly surpassed, their design must continuously evolve to keep up with model advancements.

3.3.3 Live Benchmarks

Building on the concept of highly curated benchmarks like Humanity's Last Exam [146], which rely on newly collected private data, live benchmarks offer an alternative approach to addressing data leakage issues, where test data or closely related data points appear in training sets [28]. This concern is further heightened when datasets are combined without adequate documentation of their origins [39]. As a new trend for this issue, researchers are increasingly turning to live benchmarks that continuously update with new data from data streams. For instance, LiveXiv [172] retrieves domain-specific manuscripts from arXiv in real time to automatically generate VQA pairs. While this general approach inherently mitigates contamination by constantly refreshing the dataset, it can also incur high costs, particularly when benchmark generation relies on domain experts or human evaluators and annotators.

3.3.4 Task Complexity Represented in Benchmarks

Over time, both task complexity and model sophistication have increased. For instance, early benchmarks focused on simple tasks like classifying handwritten single digits in 1988 [100], whereas more recent benchmarks involve solving complex math problems from 2021 [71] or even olympic-level problems [76]. This evolution reflects not only the advances in LLMs but also a shift towards the use of autonomous Agents.

Unlike LLMs, Agents initiate their process by receiving a user input. This normally is a direct command or a dialogue engagement from a human user. Once the task is clearly defined, they independently plan, execute the required steps, and occasionally consulting the human user for additional information or judgment. It should be noted, that *Agent* is a rather loosely defined term and often refers to an agentic workflow, where steps and scaffolding are defined beforehand [221].

The trend in rising task complexity can be quantified by assessing AI performance based on the length of tasks that agents can complete. Recent findings indicate that this metric has been growing exponentially over the past six years, with the capacity doubling approximately every seven months [97]. It is important to note that the selection of tasks and the threshold for defining success significantly influence this observed trend.

3.3.5 Considering Alternative Benchmarks

Benchmarks, while essential for assessing progress in AI, require substantial effort to generate and maintain [64]. With the increasing shift toward agent-based systems and the accelerated pace at which existing benchmarks are surpassed, researchers are now exploring strategies to derive high-quality benchmarks from existing resources. One promising direction is the use of video games, which inherently incorporate rising levels of complexity, making them ideal candidates for challenging and dynamic validation tasks for agents [143].

The utilization of video games as benchmarks in ML benchmarking is well established. Notable examples include AlphaGo [177], which combined deep reinforcement learning with Monte Carlo tree search, and AlphaStar [195], which leveraged multi-agent reinforcement learning to surpass human-level performance in StarCraft II. Early research in this area focused on reinforcement learning models tailored to single specific tasks within these games. However, there is now a shift towards utilizing generalist models capable of advancing across a variety of games. A recent illustration of this trend is found in the validation of Pokémon Red during the Sonnet

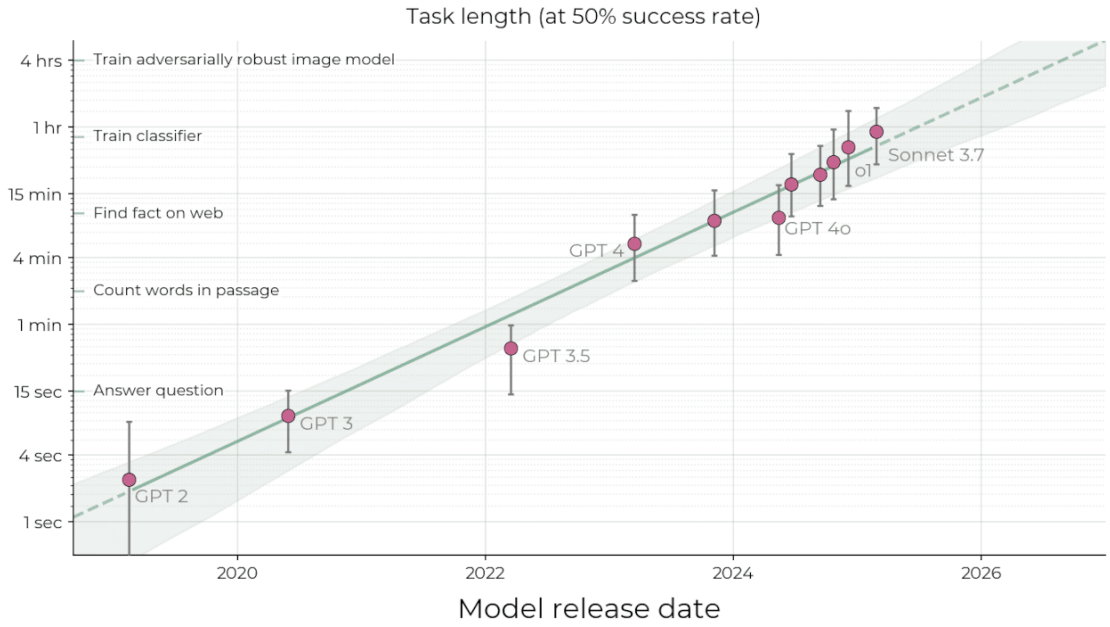


Figure 3.4: **AI's task length doubles every seven months.** Measured as the length of tasks that agents can complete at 50% success rate. The figure is adapted from [97] (license: CC BY).

3.7 assessments [183]. The single Agent was equipped with scaffolding such as basic memory, screen pixel input, and function calls for navigation and button pressing, enabling it to engage with the game beyond conventional context limitations. Other notable examples move towards benchmarking multi-agent setups, where several agents collaborate at the same time in visually-rich environments with dynamic interaction in Minecraft [112]. Minecraft is a sandbox video game that features a procedurally generated, block-based environment with intrinsic complexity and open-ended dynamics. It can be expected that future agents will be evaluated on a broad range of video games with varying problem types.

In summary, benchmarks can be constructed from a broad range of data sources. Instead of recording new data, users can save a lot time by creating a new benchmark from existing data. In the end, the data should be representative for the capabilities the benchmark creator want to measure.

3.3.6 Task Augmentation in Benchmarks

Task augmentation refers to *"generating multiple diverse tasks from a single existing task"* [131]. Prior work can be grouped into frameworks that enable the generation of tasks and datasets that were constructed through task augmentation. The method overview is presented in Table 3.1. A substantial body of research concerns large datasets created in this way, such as Visual Genome [93] and the Open Images Dataset [96]. These datasets typically provide scene graphs or large sets of annotations, allowing users to filter the data according to their specifications. Generating new tasks through task augmentation generally yields tasks that align more closely with user requirements, and existing methods cover a diversity of core perception tasks.

Criteria	CLEVR	Task Me Anything	Taskonomy	Wang 2023	JourneyBench	ProVision
	[82]	[223]	[215]	[199]	[201]	[222]
Real images/objects	❌	☑️ (partly)	✅	✅	❌	✅
Diversity core perception tasks	✅	✅	✅	✅	✅	✅
Focus on resource efficiency	Synthetic data	Strong synthetic data focus and flexible	Not applicable	❌	❌ (2,200 h of human annotation)	✅
Not reliant on generative models	✅	✅	✅	❌	❌	❌
Object-centric	✅	✅	❌	✅	❌	✅
Validated across multiple visual content domains	❌	❌	❌	✅	✅	✅
Human Ambiguity Scores	❌	❌	❌	❌	❌	❌
Easily scalable	✅	✅	❌	✅	✅	✅
Task creation code public	✅	✅	❌	❌	✅	✅
Evaluated on SOTA VLMs	❌	☑️(limited # of proprietary models)	❌	☑️ (but SOTA* 2023)	☑️	❌

Table 3.1: **Comparison of task augmentation methods.** The table compares synthetic and real-image task augmentation methods along multiple methodological criteria, such as image type, extensibility and scalability. It highlights systematic differences in design trade-offs across existing frameworks, showing how certain methods prioritize generative models and scale, whereas others prioritize adaptability to new domains. While all methods present a diversity of core perception tasks, none of the methods includes human ambiguity scores for the annotated data. Table adapted from own work [156].

At the same time, none of these methods incorporate human ambiguity scores for the underlying annotations to the best of our knowledge, instead assuming consistent annotation quality in the base real-world image data. This assumption propagates errors from the original annotations into the generated tasks, which is particularly problematic in domains with high inter-rater variability, such as medicine, because that undermines the reliability on downstream tasks. Moreover, evaluations are usually limited to a small set of state-of-the-art (SOTA) VLMs, further restricting interpretability.

In contrast to purely rule-based or algorithmic techniques, the integration of generative models has made it possible to scale task generation pipelines more rapidly. This shift introduces new limitations, including hallucinations [200], reduced controllability compared to rule-based methods, and undesired artifacts such as spurious correlations.

In summary, we observe a growing number of task augmentation methods, particularly those that rely on generative models, though substantial methodological challenges remain.

3.3.7 Quality Assurance in Benchmarks

As previously highlighted, the number of new benchmarks is rapidly increasing. While benchmarks aim to reliably identify the best performing methods and expose their strengths and weaknesses, the quality of individual data points within benchmarks warrants careful consideration. Recent research across various domains, ranging from code generation benchmarks [7] to broad STEM

benchmarks [193, 7, 38], indicates that label noise stemming from ambiguous or outright mislabeled data point can significantly distort the interpretation of benchmark results [193, 7, 38]. Domains, such as biomedical or surgery [117] are known for their high ambiguity level that is quantified with inter-rater variability that can be reduced by providing further context to the annotators [211]. For these domains, the measured ambiguity reflects the upper ceiling of data quality. However in the reported cases from non medical domains [193, 7, 38], the inter-rater variability is zero or minimal. Consider the following question from the Simple Variations on Arithmetic Math word Problems (SVAMP) benchmark, whose original answer is objectively incorrect. *"Question: You had 14 bags with equal number of cookies. If you had 28 cookies and 86 candies in total, how many bags of cookies do you have? Original Answer: 2. Correct Answer: 14*" [193].

Consequently, the updated versions removed error that otherwise would undermine the reliability of the benchmark, potentially leading to misleading rankings or erroneous conclusions about the best performing model. Additionally, benchmarks that build upon widely adopted task standards, such as the UK AISI INSPECT standard [4] tend to achieve greater adoption by the community since they reduce implementation complexity for new users.

In summary, we observe a growing trend of benchmark creators revisiting established datasets to release improved iterations, addressing annotation ambiguities, refining task descriptions, and harmonizing implementation details. Such iterative refinement enhances benchmark reliability, lowers barriers for adoption, and contributes to more robust and informative comparisons among AI models.

Link to this thesis (RQ2)

The rapid developments and limitations of existing benchmarks highlight the need for user-centric benchmarking approaches. This motivates our investigations on scaling user-centric benchmarks with DomainBench (**RQ2**) and its application to the high-stakes domain of surgical imaging (**RQ2**).

4 | Contributions

In this chapter, we outline the four primary contributions of this thesis, encompassing the areas of data annotation (see Section 4.1), internal QA (see Section 4.2), domain benchmarking (see Section 4.3), and lastly its application to surgical imaging data (see Section 4.4). For each contribution, we highlight my contributions to the presented work.

4.1 Quantifying the Impact of Labeling Instructions

Author's Note on Reused Material:
This section and the referenced appendix sections are based in large parts on the peer-reviewed publication: Tim Rädsch et al., "Labelling Instructions Matter in Biomedical Image Analysis", published in *Nature Machine Intelligence* in 2023 [155]. As the lead and first author of this publication, I played the central role in designing the study, preparing and conducting the annotation experiments, and implementing the associated code. I led the development and execution of the competition analysis and designed all figures. I initiated and coordinated the collaboration with the involved annotation companies and annotators and secured the funding necessary for the annotation work. I led the interpretation of results and led the writing of the manuscript together with several co-authors, who provided analytical input, textual contributions, and iterative feedback. Significant portions of the text, figures, and results have been adapted and integrated into this chapter. Permissions for text and figure reuse have been confirmed (license: CC-BY-4.0). The concept passage has been rewritten so that each experiment's rationale, finding, and implication is clearly explained, and the respective Discussion subsection enables readers to understand both the integrated contribution and its individual ramifications. This work was co-authored with Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Schreck, A. Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein.

4.1.1 Concept Overview

The validation of biomedical image analysis algorithms critically relies on the availability of high-quality annotated reference datasets. A central foundation of annotation quality in such datasets is the specification of clear and effective labeling instructions. Despite their pivotal

role, the systematic development and empirical evaluation of labeling instruction quality have received remarkably little attention in the biomedical imaging domain.

This chapter presents a structured investigation into the design of labeling instructions and their influence on annotation quality. Drawing upon a multifaceted analysis that encompasses both a review of annotation practices in international biomedical image analysis competitions (as registered by the MICCAI Society) and a large-scale annotation experiment, we identify a pronounced gap between the actual needs of annotators and the quality and availability of instructions typically provided in practice. The overview is displayed in Figure 4.1.

The experiment comprises annotations of 14,040 biomedical images by 156 annotators from four professional annotation companies and 708 crowdworkers from Amazon Mechanical Turk (MTurk), each operating under systematically varied instruction conditions. We assess the impact of instruction information density on annotation performance with varying instructions ranging from minimal text to rich text, and accompanied by visual examples. The results demonstrate that the inclusion of exemplary images significantly improves annotation quality relative to text-only guidance, whereas increased textual detail alone does not yield similar gains. Moreover, professional annotators consistently produce higher-quality segmentations than MTurk contributors across all instruction types.

Taken together, this work provides the first empirical foundation for establishing best practices in the formulation of labeling instructions within biomedical image analysis. It highlights the urgent need for standardized, example-driven instructional design to ensure reliable reference data for algorithm validation.

4.1.2 Methods

Following the definition of terms used throughout this study, we will describe the selected data, the annotation providers, the labeling instructions, the experimental setup, and the statistical analysis in detail throughout this section.

Definitions

We use the following terms throughout this contribution:

Annotation provider: An entity that provides annotations performed by human workers. They can be categorized into two types of annotation providers: a) crowdsourcing platforms, such as MTurk, and b) professional annotation companies (see the following definitions).

Annotation requester: An entity that wants a dataset annotated by an external annotation provider.

Challenge/Competition: We follow the *challenge* definition of the BIAS statement, which defines a challenge as an "open competition on a dedicated scientific problem in the field of biomedical image analysis. A challenge is typically organized by a consortium that issues a dedicated call for participation. A challenge may deal with multiple different tasks for which separate assessment results are provided. For example, a challenge may target the problem of segmentation of human organs in Computed Tomography (CT) images. It may include several tasks corresponding to the different organs of interest." [114]

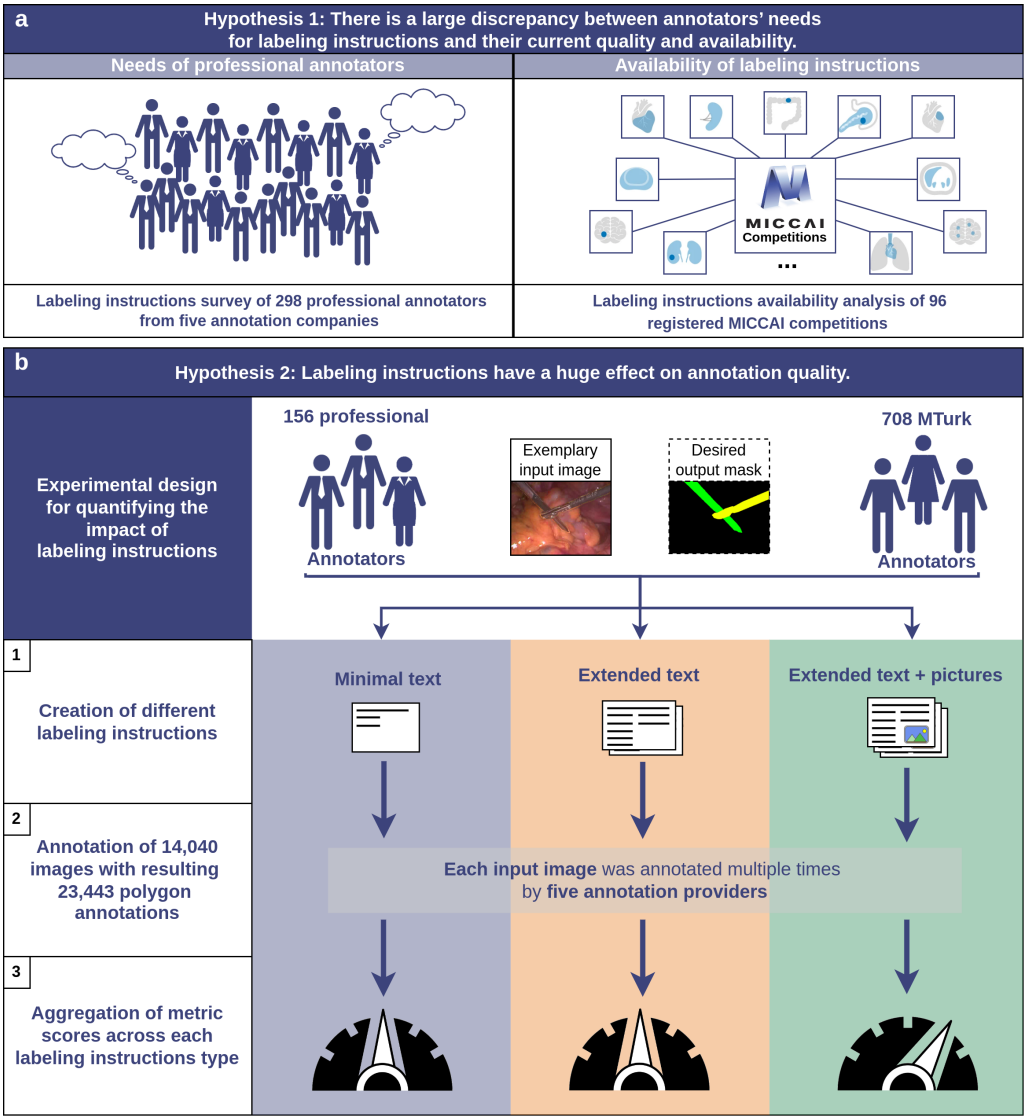


Figure 4.1: **Hypotheses of this contribution and overview of methodology.** (**a**) Annotators' needs for labeling instructions (left side) were captured via a survey. The availability of labeling instructions (right side) was captured via international competitions conducted in the scope of the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) conference. (**b**) To assess the impact of labeling instructions on annotation quality, three types of labeling instructions were created for the same dataset: Instructions using i) *minimal text*, ii) *extended text* and iii) *extended text including pictures* were issued to a total of 864 annotators from five different labeling providers. The resulting dataset of 14,040 annotated images was analyzed with a two-part beta mixed model. The dials qualitatively represent the observed annotation performance.

Challenge/Competition task: The BIAS statement defines a *challenge task* as a "subproblem to be solved in the scope of a challenge for which a dedicated ranking/leaderboard is provided (if any)." [114]

Labeling instruction: A document or tool that specifies how to correctly label (imaging) data. The different types of labeling instructions are defined below and presented in greater detail in Appendix A.5-A.7.

MTurk: A two-sided crowdsourcing marketplace that enables annotation requesters to hire freelance workers, referred to as MTurk crowdworkers, to perform discrete tasks on-demand.

MTurk crowdworker: A remotely located person performing discrete on-demand tasks on MTurk on their own hardware. They are employed on a freelance basis.

Professional annotation company: A company focusing mainly on generating annotations for (imaging) data. Their workers are located in regular office space and mainly employed full-time.

Professional annotator: An on-site located and full-time employed person performing annotations for a professional annotation company in a provided office space with according hardware.

International Professional Annotator Survey

To obtain a comprehensive understanding of the current issues professional annotators face with respect to labeling instructions and their work characteristics, we developed a 26-item questionnaire (provided in Supplementary Information 4). The survey was distributed among five internationally operating annotation companies in best-cost countries that exclusively employ professional annotators. To increase the statistical validity of the submitted entries (n = 363), we employed a twofold filtering strategy of the entries: a) a control question pair, which consists of the positive and negative formulation of a question, and b) an instructional manipulation check, as recommended by Oppenheimer et al. [139]. The check asks the participant to answer an open-ended question with a specific set of words that can only be answered by carefully reading the question text. The filtering resulted in 298 remaining entries.

Competition Analysis

Our goal was to capture the current handling of labeling instructions in biomedical image analysis. Thus, we included all Medical Image Computing and Computer Assisted Intervention Society (MICCAI) registered competitions which were published until the end of 2021. We retrieved the registered competitions from the MICCAI website, where all MICCAI registered competitions are published. This resulted in a list of 53 competitions with 96 competition tasks. Two engineers with a proven history in reviewing competitions rated all submitted standardized competition design documents of the individual competition tasks as to whether a labeling instruction was provided by the competition task. In addition, ambiguous cases were marked as such. This resulted in an inter-rater agreement of 90.6%. Contradictory ratings were mainly cases where both raters marked the competition task as ambiguous and were solved by an independent third engineer with a proven history in reviewing competitions. 20 competition tasks were excluded as not applicable in the process, as they provided a valid reasoning why labeling instructions cannot be provided. An example is the Medical Out-of-Distribution Analysis Challenge [226], where the publication of the labeling instruction would enable cheating, because it contains

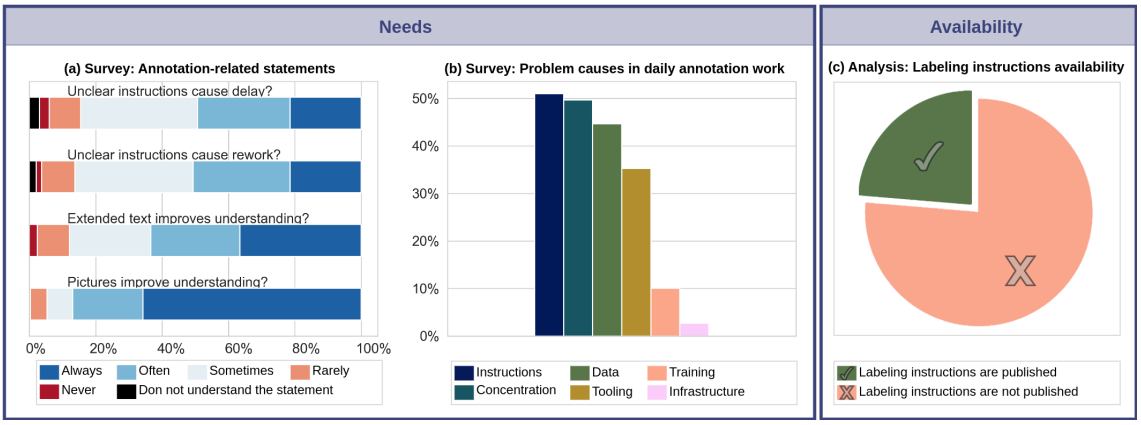


Figure 4.2: **The field of biomedical image analysis suffers from a notable discrepancy between the needs of those annotating the data and the actual availability of labeling instructions (if any).** (**a**) Professional annotators agree that unclear instructions consistently cause delay and rework. Comprehensive text descriptions and images are perceived to improve annotation quality. (**b**) Professional annotators attribute labeling instructions as the primary cause for problems related to their daily annotation work, followed by concentration issues and poor input data. Answers for (a) and (b) were processed from 298 annotators from five annotation companies. (**c**) 76% of the recent Medical Image Computing and Computer Assisted Intervention Society (MICCAI) conference competition tasks, matching the inclusion criteria, do not report any labeling instructions. The analysis includes all 96 registered MICCAI competition tasks between 2020 and 2021.

information about the placement of out-of-distribution objects in the competition data. The result of the competition tasks analysis is provided in Supplementary Information 3.

Dataset Selection

The Heidelberg Colorectal (HeiCo) segmentation [116, 166] dataset, comprising medical instrument segmentations in laparoscopic video data, served as the basis for this study. Each image was enhanced by 21 meta annotations representing relevant image characteristics or artifacts (e.g. whether overlapping instruments or motion blur are present on the image), where each image characteristic is a binary annotation decision [166]. The meta annotations were implemented by a trained engineer with extensive annotation experience, who was involved in the original creation process of the dataset. Based on the meta annotations, the images were categorized into nine different image categories, which represented the potential annotation difficulty and served for the subsequent image selection:

1) Simple category: Images do not contain any artifact on the instruments.

2) Chaos category: Images contain at least three different artifacts on the instruments. Moreover, images containing a higher number of instruments are preferred.

3) Trocar category: At least one trocar is present on the image.

4) Intersection category: At least two medical instruments are intersecting on the image.

5) Motion blur category: A minimum of one medical instrument with the motion blur artifact is

present on the image.

6) Underexposure category: At least one medical instrument on the image is underexposed.

7) Text overlay category: Text overlay is present and obstructs the view of the image.

8) Image overlay category: An image overlay is present and obstructs the view of the image.

9) Random category: Images are randomly selected from the remaining images in the test set.

The goal was to select a broad range of image categories and annotation scenarios. Therefore, we selected 234 unique frames corresponding to the defined categories. Each unique image was annotated four times per labeling instruction and per annotation provider, resulting in 60 annotations per image (generating a total of 14,040 annotated images). 15 unique images from category 1 to category 8 were selected by hand, accounting for roughly half of the images. The only exception was category 8, because there existed only eleven unique images that matched the definition of the category. The other half was selected randomly (category 9).

Annotation Providers

The study was conducted based on five annotation providers, consisting of four professional annotation providers and the crowdsourcing platform MTurk. Each annotator participated exclusively with one of the three labeling instructions. We selected high quality representatives for the professional annotation companies and the crowdsourcing platform. The professional annotation companies operate internationally and their annotators are located in best-cost countries. The selected companies had a proven track record in large-scale industry annotation projects. This was particularly relevant, because the number of annotation providers has been growing steadily for the last ten years [31]. Participating crowdworkers on MTurk had to fulfill the following quality requirements: a) 98% accepted HITs and b) a minimum of 5,000 accepted HITs. Our quality requirements thus far surpassed the quality requirements of researchers working with MTurk, which normally require 95% accepted HITs with a minimum of 100 accepted HITs [89, 67, 15, 25]. To capture a representative sample of the population on MTurk, we spread our HITs across 40 days and all times of day. We followed Litman et al. to ensure a fair worker compensation [111]. The base annotations and initial analysis were originally collected as part of a prior publication [151].

Labeling Instructions

As a labeling instruction specifies how to correctly label the (imaging) data, we defined a set of design rules that are shared across all labeling instructions:

(a) The information of a labeling instruction is provided in a slide layout for better human information processing.

(b) Each slide represents a chunk, an encapsulated unit of information. Chunking reduces the demand on the working memory.

(c) Related information chunks are positioned near each other.

(d) A consistent layout with defined fonts, symbols and colors is applied.

Minimal text labeling instructions: The *minimal text labeling* instructions consist of a limited textual description, including positive examples and the most common annotation occurrences

(see Appendix A.5). This represents a situation where only little effort was put into creating the labeling instruction. For example, text overlay references the uncommon occurrence of text that is visible in the image. Because it is an uncommon occurrence, it is not mentioned in the *minimal text labeling* instruction. As a baseline, the *minimal text labeling* instructions consist of seven slides with 168 words.

Extended text labeling instructions: The *extended text labeling* instructions extend the *minimal text labeling* instructions with a comprehensive text description, which is supported by both positive examples and counterexamples in text form. Furthermore, both common and uncommon cases are included, see Appendix A.6. In this labeling instruction for example, the uncommon occurrence of text overlay is described in detail. This resulted in ten slides with 446 words.

Extended text including pictures labeling instructions: The *extended text including pictures labeling* instructions complement the *extended text labeling* instructions with pictures, see Appendix A.7. The pictures include textual descriptions, symbols, markings and the usage of color to convey the information on the slides. In addition, rare annotation occurrences are included as well. This represents a situation where extensive (domain) knowledge about the labeling process is present and documented in detail in the labeling instruction. Hence, in our example, the uncommon occurrence of text overlay is described in detail with text and pictures. These labeling instructions consist of 16 slides with 961 words.

4.1.3 Experiments

In the following, we describe the experimental setup for quantifying the impact of labeling instructions, and the statistical analysis.

Labeling Instructions Analysis

Each of the five labeling providers annotated the same images subsequently with each labeling instruction, using separate annotators. We started with the *minimal text labeling* instructions, followed by the *extended text labeling* instructions and provided the *extended text including pictures labeling* instructions last, to prevent information leakage. Providing the instructions in parallel or in a different order would enable participants to receive information from instructions with a higher information density. As an additional security measure, we added a minimum break of ten days between two labeling instructions to prevent any potential overlap or sharing. No individual questions from the annotators regarding the labeling instruction content were answered in order to prevent a potential information advantage for an annotation provider that could impact the statistical analysis. Each annotator was only allowed to participate for a single labeling instruction. We tracked all accounts over the duration of the study to ensure that everyone has only access to a single instruction. After completion of an annotation stage, the accounts were deactivated. All participating annotators had no prior experience with their respective annotation tooling, minimizing their impact. Furthermore, no worker selection tasks were used for the annotation providers, since annotation companies prefer to allocate experienced annotators to new projects to win potential follow-up projects. Each professional annotator annotated a total of 72 images. To properly simulate the parallelization of crowdsourcing, each MTurk crowdworker annotated four images. After the submission of their annotations, each MTurk crowdworker could submit a short optional survey about their work characteristics. The

survey enabled us to gather a broad representation of worker characteristics across different stakeholders.

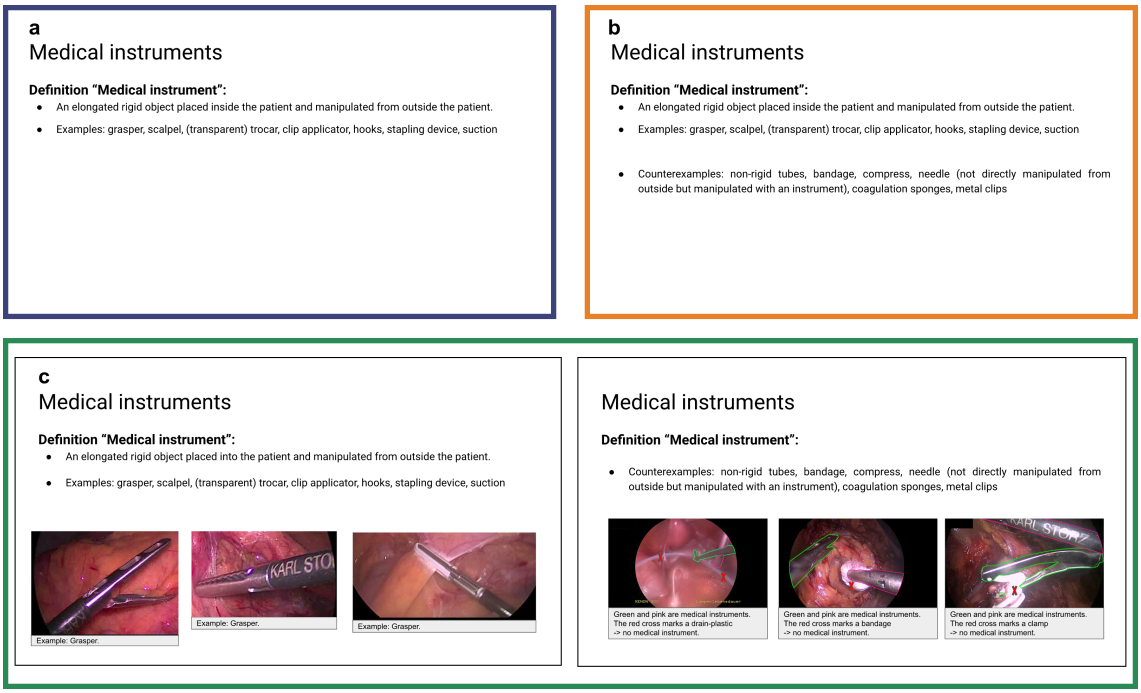


Figure 4.3: **Example of labeling instructions.** Medical instruments are initially defined in the *minimal text labeling* instructions (**a**). The definition is deepened in the *extended text labeling* instructions (**b**) and enriched with images in the *extended text including pictures labeling* instructions (**c**).

Statistical Analysis

To quantify the impact of labeling instructions and the two annotator types, the following statistical methods were used:

To ensure compatibility with prior work on the data, we utilized the same metrics as suggested by Roß et al. [166], in which the dataset was originally introduced as part of the MICCAI Robust Medical Instrument Segmentation Challenge 2019.

$$\mathrm{DSC}(A, B) = \frac{2\,|A \cap B|}{|A| + |B|}, \tag{4.1}$$

where A is the set of predicted pixels, B is the set of reference pixels, and $|\cdot|$ denotes the cardinality (here: area).

$$\mathrm{NSD}^{(\tau)}(A, B) = \frac{|S_A \cap \mathcal{B}_B^{(\tau)}| + |S_B \cap \mathcal{B}_A^{(\tau)}|}{|S_A| + |S_B|}, \tag{4.2}$$

where S_A and S_B are the boundary pixels of A and B, $\mathcal{B}_A^{(\tau)}$ and $\mathcal{B}_B^{(\tau)}$ denote the border regions within distance τ from S_A and S_B respectively, and $|\cdot|$ denotes the cardinality of the set.

We analyzed the annotation results based on the DSC scores with a two-part zero-inflated beta mixed model [21]. Missing an annotation and providing a subpar annotation are fundamentally different error types. There we differentiate the error types in our statistical model as follows: a) the first part included a logistic mixed model analyzing the probability of severe annotation errors (at least one instrument with DSC = 0 in one frame) and b) the second part consisted of a beta mixed model analyzing the non-zero DSC values of an image when valid annotations occurred. The image variable and the annotation worker variable were modeled as random effects while the type of labeling instructions, annotator type, image category and access to context video were modeled as fixed effects with two-side hypothesis tests. The model was implemented in the brms package in R [185], where vague Gaussian priors centered on 0 were used for the fixed effects and Half-Cauchy priors were assigned for the standard deviation of the random effects. 4,000 Markov Chain Monte Carlo samples were generated across 4 chains. The obtained estimates of the covariates are on the log-odds scale and were exponentiated to obtain the odds ratio for each covariate. Software: R version 4.0.2 (package brms version 2.16.0). The implementation was conducted by an experienced statistician to ensure that the selected analysis and implementation adhere to the highest standards.

$$\mathsf{I} \sim (1 \mid \mathsf{Worker\_ID}) + (1 \mid \mathsf{Image\_name}) + \sum_{i=1}^{4} F_i \tag{4.3}$$

where:

- I is the response variable. It equals 0 if Image_min_instance_DSC_value $= 0$, and equals the image DSC score otherwise.
- $(1 \mid \text{Worker\_ID})$ represents a random intercept for each Worker_ID.
- $(1 \mid \text{Image\_name})$ represents a random intercept for each Image_name.
- $\sum_{i=1}^{4} F_i$ represents the fixed effects, with F_i being the i^{th} analyzed covariate.
- Both the mean and zero-inflation components of the model include the same fixed and random effects.
- The distribution is zero-inflated beta with regression coefficients assigned $\mathcal{N}(0, 10)$ priors and random-effect standard deviations assigned Cauchy$(0, 1)$ priors.

List of covariates:

- Fixed effects:
  - Stage
  - Annotator type
  - Image category
  - Context video
- Random effects:

  - Worker_ID
  - Image_name
- Image category levels:
  - Randomly selected images
  - Simple cases
  - Chaos cases with multiple artifacts
  - Images with overlays
  - Intersecting objects
  - Images with motion blur
  - Images with text overlay
  - Images containing a trocar
  - Underexposed images

4.1.4 Results

Given the lack of (1) awareness of the importance of labeling instructions and (2) quantitative research investigating how to best perform the labeling, we initiated our study by systematically analyzing the perspective and work characteristics of professional annotators, and common practice of labeling instructions in leading biomedical imaging competitions. Subsequently, we investigated the impact of labeling instructions with varying levels of information density on the annotation quality, and the effect of different annotator types on biomedical imaging data. The information levels included instruction with minimal text, extended text or extended text including visual examples.

Professional Annotators Request better Labeling Instructions

To motivate our empirical study on annotation quality, we conducted an international survey among 363 (298 after filtering noisy answers) professional annotators employed by five different internationally operating annotation companies. We selected companies that had completed large-scale annotation projects in the past in the automotive sector. Depicted in Figure 4.2a-b, the results reveal that the majority of annotators request more time and resources to be spent in the generation of labeling instructions. In fact, poor labeling instructions were identified as the primary cause of problems related to annotation work followed by concentration issues (50%) and poor input data (45%).

The Importance of Labeling Instructions May Be Underrated

Despite their apparent importance, earlier research revealed that labeling instructions are typically not provided and/or reported in the field of biomedical image analysis [114]. This even holds true for international image analysis competitions although these can be expected to provide particularly high quality with respect to validation standards. To address this issue, the MICCAI, the largest international society in the field, took action and developed a comprehensive reporting guideline [115] for biomedical image analysis competitions. The guideline comprises

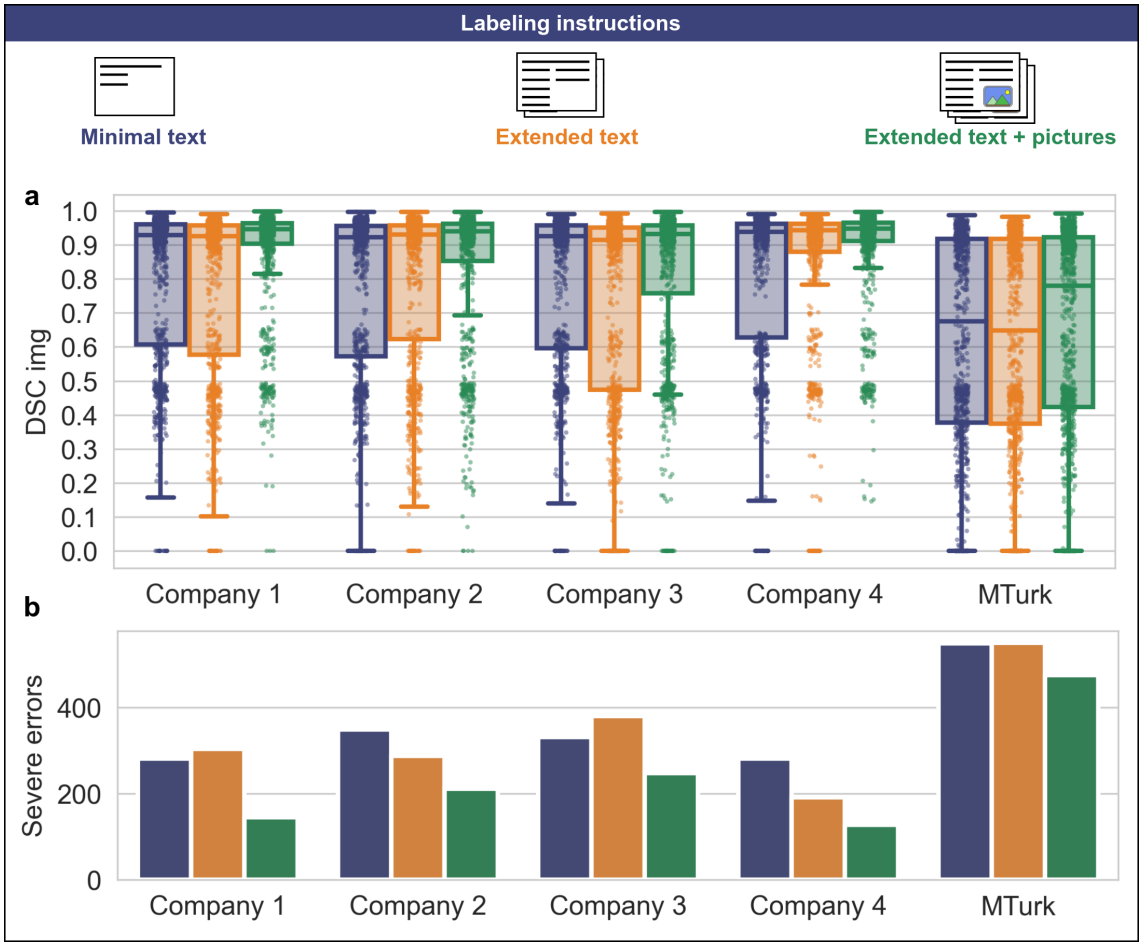


Figure 4.4: **Key findings of this contribution:** (1) Extended text descriptions (orange) do not necessarily boost annotation performance compared to minimal text descriptions (blue), while (2) including images (green) gives a clear benefit for all annotation providers. (3) Professional labeling companies (Companies 1 - 4) provide substantially higher-quality annotations compared to the most popular crowdsourcing platform Amazon Mechanical Turk (MTurk). **(a)** The Dice Similarity Coefficient (DSC) score has been aggregated for each annotated image and is displayed aggregated for each pair of company and labeling instruction as a dots- and boxplot (the band indicates the median, the box indicates the first (25th percentile) and third (75th percentile) quartiles and the whiskers indicate $\pm 1.5\times$ interquartile range, the DSC score maximum is 1 and the minimum is 0 for each image). **(b)** The absolute number of severe annotation errors, defined as annotations with a metric score equal to zero, is also shown. A total of 14,040 images were annotated by 156 annotators from four professional companies and 708 MTurk crowdworkers.

an entire paragraph on reporting the annotation process, including the labeling instructions. Prior to conducting a competition in the scope of a MICCAI conference, researchers must put the report for their competition online [124] to foster transparency and reproducibility, and to prevent cheating [160]. To capture the SOTA regarding labeling instructions in biomedical image analysis, we analyzed all MICCAI competitions officially registered in the past two years (PRISMA statement [142]: Appendix A.1). Although the reporting guideline explicitly asks for (a link to) the labeling instructions, 76% of the recent MICCAI competitions do not report any labeling instructions (cf. Figure 4.2c). Given that MICCAI competitions make up around 50% of the biomedical image analysis competitions in a year [114], this can be regarded as a widely spread phenomenon.

In current biomedical image analysis practice, labeling instructions are thus often neither of sufficient quality, nor are they appropriately reported and valued in the scientific community. Both issues negatively impact scientific quality in the field and hinder reproducibility.

Extended Text Descriptions Do Not Boost Annotation Performance

The shortcomings in quality we found in common practice regarding labeling instructions called for an investigation on how this quality can be improved. As a first step in this direction, we sought to determine the impact of different types of labeling instructions on the quality of annotation of a particular dataset. The selected dataset [116, 166], which can be handled by crowdworkers, combines highest quality reference annotations and 21 meta annotations per annotated image that reflect the annotation difficulty. We created three distinct types of labeling instructions with varying levels of information density, namely (a) *minimal text*, (b) *extended text*, and (c) *extended text including pictures*, as detailed in Methods and Appendix A.5-A.7. Example instructions are provided in Figure 4.3. The obtained 23,443 annotations on 14,040 images from the 864 annotators were analyzed with a two-part zero-inflated beta mixed model (ZIBMM).

In contrast to the *limited text labeling* instructions, the *extended text labeling* instructions included more detailed descriptions, counterexamples and information on uncommon annotation cases that might appear. We define an annotation with a metric score equal to zero as a severe annotation error. For the 4,680 images annotated with the *extended text labeling* instructions, we observed a minor increase in the number of severe annotation errors compared to the *limited text labeling* instructions [median: +0.4%; max: +14.8%; min: -31.7%]. Furthermore, we observed no impact on the median DSC [43], and only a minor increase in the Interquartile Range (IQR) [median: +1.8%; max: +33.3%; min: -75.8%]. These results contradict the initial assessment of the professional annotators. The absent effect of the *extended text labeling* instructions is reinforced by the results of the two-part ZIBMM, where we obtained no statistically significant difference for the *extended text labeling* instructions compared to minimal text from both the first and second part of the model, revealing that extended text descriptions do not boost annotation performance.

Exemplary Images Are Crucial for Annotation Quality

Professional annotators claim that pictures help them understand labeling instructions (cf. Figure 4.2a). Therefore, the *extended text including pictures labeling* instructions were enriched by pictures including rare occurrences. In comparison to the *extended text labeling* instructions, the number of severe annotation errors was reduced for all five annotation providers [median: -33.9%; max: -13.6%; min: -52.3%]. Furthermore, their median DSC score increased [median:

+2.2%; max: 20.0%; min: +1.1%], and their IQR was reduced [median: -58.3% ; max: -9.1%; min: -84.2%] (cf. Figure 4.4a). This reinforces professional annotators' initial assessment that pictures improve their understanding (cf. Figure 4.2a). The improvements occurred mainly on the difficult annotation cases (cf. Figure 4.5). Surprisingly, the extended text instruction increased the number of lower Based on the two-part ZIBMM, the odds of obtaining a severe annotation error with these labeling instructions are 0.37 times (Credible Interval (CI): 0.28, 0.50) that with *minimal text labeling* instructions. From the second part of the two-part ZIBMM, we obtained no significant difference in the DSC score, once an object was identified. Thus, the improvements primarily stemmed from the additional reduction of severe annotation errors (cf. Figure 4.4b).

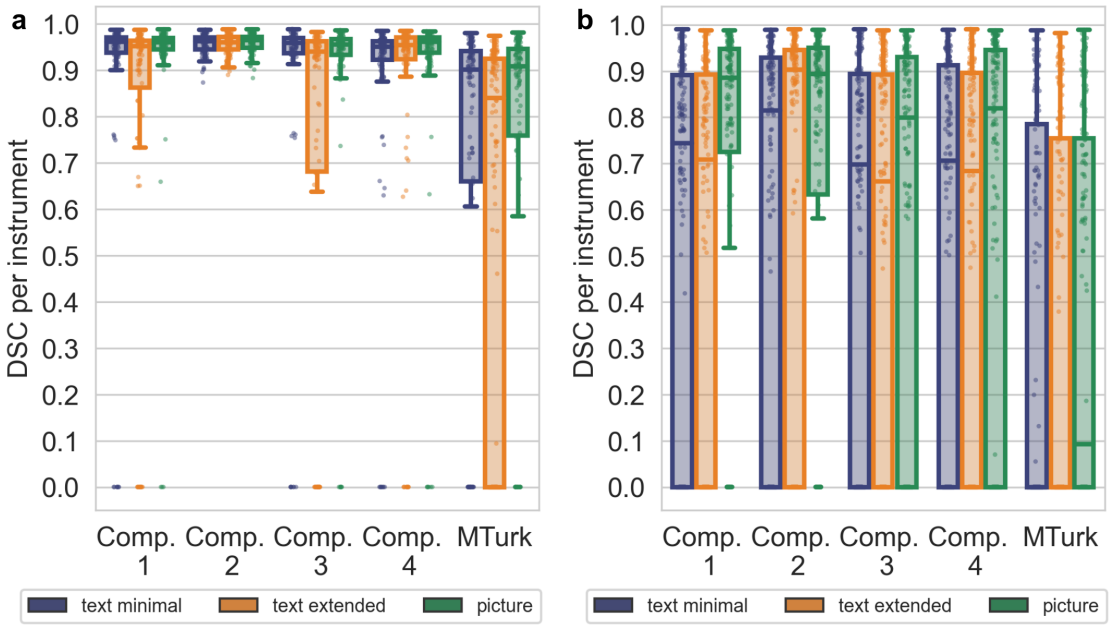


Figure 4.5: **DSC scores for the category with the least and the most image characteristics present on the instrument level.** For each annotation provider and type of labeling instruction, the Dice Similarity Coefficient (DSC) scores per instrument are aggregated. (**a**) displays the category with images that do not contain any artifact on the instruments. (**b**) displays the category with images that contain at least three different artifacts on the instruments. The DSC scores for (a) and (b) are displayed as a dots- and boxplot (the band indicates the median, the box indicates the first (25th percentile) and third (75th percentile) quartiles and the whiskers indicate $\pm 1.5\times$ interquartile range, the DSC score maximum is 1 and the minimum is 0 for each image). A combined total of 3,771 instrument annotations were observed for the category with the least and the most present image characteristics.

Professional Labeling Companies Outperform The Most Popular Crowdsourcing Platform

In comparison to MTurk crowdworkers, professional annotators conduct labeling as their main source of income, label more often in a week, and label for a higher number of weekly hours (cf. Figure 4.6a - c). In contrast, MTurk workers have a longer employment history in labeling than professional annotators, as displayed in detail in Figure 4.6d. This observation is consistent with the historical development of the data annotation market, where general labor markets, including

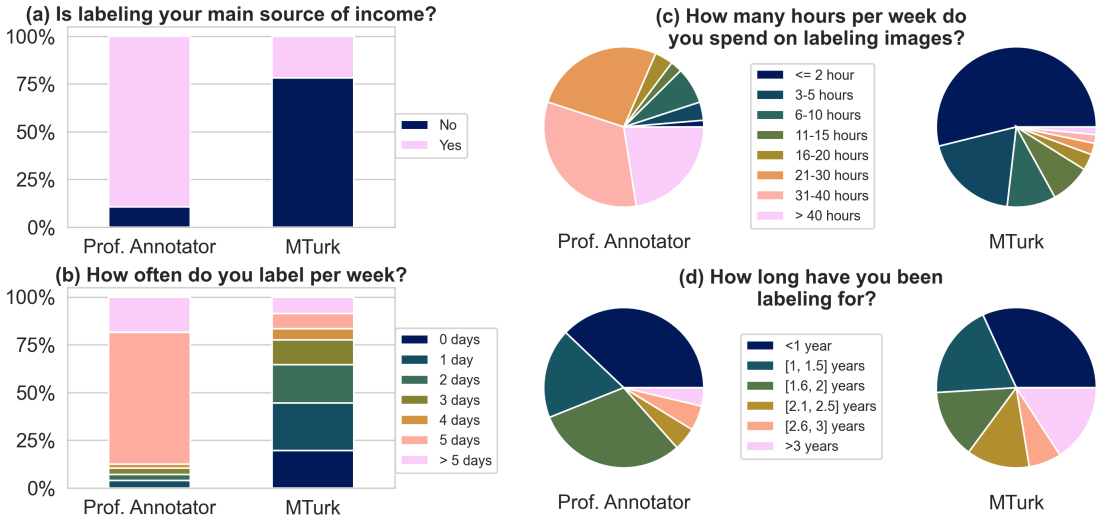


Figure 4.6: **Work characteristics of professional annotators and Amazon Mechanical Turk (MTurk) crowdworkers.** In comparison to the MTurk crowdworkers, professional annotators conduct labeling as their main source of income (**a**), label more often in a week (**b**), and spend a higher number of hours per week on labeling images (**c**). In contrast, MTurk crowdworkers have a longer employment history with labeling than professional annotators (**d**). Answers were processed from 298 professional annotators from five annotation companies and 518 MTurk crowdworkers.

MTurk, preceded professional companies. Given *minimal text labeling* instructions, professional annotators produced less severe annotation errors [companies median: 307; MTurk median: 549]. Furthermore, the professional annotators' annotations generated a higher median DSC score [companies median: 0.93; MTurk median: 0.67], and a smaller IQR of the DSC score [companies median: 0.36; MTurk median: 0.67]. Both annotator types only showed minor or no improvement with the *extended text labeling* instructions. While both annotator types benefitted from added pictures, professional annotators displayed a stronger reduction of severe annotation errors [companies median: -34.4%; MTurk median; -13.6%] and of the IQR [companies median: -63.00%; MTurk median: -9.1%]. In contrast, MTurk crowdworkers displayed a stronger improvement of the median DSC score [companies median: +1.7%; MTurk median: +20.0%]. Under the same conditions, the odds of severe annotation errors for a professional annotator were 0.09 times (CI: 0.06, 0.12) that of a MTurk crowdworker, keeping all other factors constant. Similarly, once an object was identified, the odds for a professional annotator of achieving a perfect DSC score were 94.7% (1.947, CI: 1.69, 2.24) higher than those of a MTurk crowdworker.

4.1.5 Discussion

To our knowledge, this study is the first to quantitatively and critically examine the role of labeling instructions in professional and crowdsourced annotation work. We were able to uncover a major discrepancy between their importance and quality/availability, and determine which type of labeling instructions is the most effective. While labeling instructions play a crucial role in the creation of biomedical image analysis datasets, current common practice is insufficient and neither meets the requirements of industry (e.g. in creating large-scale datasets), nor those

of academia (e.g. in hosting competitions) (cf. Figure 4.2c). Notably, professional annotators demand a higher time and resource commitment from the creators of labeling instructions. Furthermore, they identify current labeling instructions as a main cause for annotation delay and rework (cf. Figure 4.2a). Interestingly, we found that extended text descriptions do not necessarily boost annotation performance compared to minimal text descriptions (cf. Figure 4.4), although professional annotators expect them to improve the understanding of the labeling tasks (cf. Figure 4.2a). In contrast, the addition of pictures resulted in a clear improvement among all annotation providers (cf. Figure 4.4), which matches the assessment of the professional annotators (cf. Figure 4.2a). This improvement was mainly observed on ambiguous images with challenging conditions, such as poor illumination or intersecting objects, as depicted in Figure 4.5. Since ML models fail their prediction more often on ambiguous images than on clear images [173], the correct annotation of such images in training datasets is particularly crucial for good performance. Lastly, annotators from professional companies provide substantially higher-quality annotations compared to those from the most popular crowdsourcing platform in health research, MTurk, regardless of the type of labeling instruction.

One of the key implications of our findings is that there is a huge discrepancy between the potential impact of labeling instructions on generating the desired annotations and their role in current (research) practice. Our study shows that scope and specific design choices of labeling instructions play a major role in generating the desired annotations. Among other contributing factors such as prior annotation expertise or training, labeling instructions may thus determine the attention to detail that annotators pay to annotating challenging images. The importance of this is amplified in cases of datasets becoming increasingly complex and diversified over time, where more and more challenging data points are added due to initial prediction difficulties of the ML model. Consequently, researchers and practitioners alike should ensure proper representation of the necessary information in their labeling instructions, extend them if needed, and invest the necessary time to produce informative annotated images. However, we observed that 76% of the recent MICCAI competitions did not report their labeling instructions (cf. Figure 4.2c). Since competitions are aimed at generating high publicity, it can be assumed that their current handling of labeling instructions represents the upper bound of quality regarding common practice, with the quality being significantly lower for research projects devoid of intense public scrutiny. This discrepancy is alarming and calls for a paradigm shift in common practice. Given our results, the MICCAI Special Interest Group [119] for Biomedical Image Analysis Challenges is currently reevaluating common competition practices and considering stricter rules for data quality for future challenges.

Another implication of our study is that generating and publishing labeling instructions is a precondition for enabling independent verification and reproduction of the created annotations. Similarly to how published code enables the verification of algorithmic results in research papers, access to labeling instructions is necessary to understand annotators' decisions and potentially recreate annotations. Furthermore, the information provided in labeling instructions is important for ML practitioners. Based on the annotation decisions (e.g. handling of occluded objects of interest) implemented in a dataset, ML practitioners need to define their own desired outcome for these occurrences and modify their ML model accordingly. Given their impact on the resulting annotations and the current poor state of datasets [135, 152, 144, 145], we argue that dataset creators and competition organizers should publish their labeling instructions, as proposed by Maier-Hein et al. [115]. To advance current practice, we recommend the dataset and labeling instruction creation to be an iterative process properly modeling the underlying distribution

of the data space, as described in Freeman et al. [51]. In earlier stages of the dataset creation process, the focus should be on common occurrences (e.g. common surgical instruments in the case of laparoscopic surgery) to generate strong initial model performance. Throughout the process, special, conflicting or rare occurrences should be added to the dataset to maximize the model performance and reflect the real-world distribution.

A further recommendation motivated by our study is for annotation requesters to evaluate their annotator options more carefully and select their provider based on suitability to their annotation requirements. While medical personnel alone may be too sparse and costly to satisfy the rising demand for annotated biomedical image data [51], oftentimes, medical domain knowledge may only be necessary for the creation of the labeling instructions and not the annotation itself. Thus, crowdsourcing in combination with computer assisted annotation strategies can be a valid and cost-effective approach. MTurk, the most commonly used crowdsourcing platform in health research [32], follows a do-it-yourself model, where all components from annotator training to annotation tooling are provided by the annotation requester and the crowdworkers are employed on a freelance basis. In contrast, professional annotation companies assign a dedicated contact person that oversees the project, with annotators trained and directly employed by the annotation company. While MTurk can quickly scale up with a large number of people, professional annotation companies tend to scale up more slowly. However, professional annotators work on data annotation for a longer proportion of their workday and are usually assigned full-time to an annotation project (cf. Figure 4.6). Regarding location, the international annotation market leads to scenarios where requesters and annotators do not share the same (native) language. This reinforces the need for clear and concise labeling instructions with exemplary pictures as a fundamental requirement for scaling data annotation operations.

Irrespective of the person chosen to conduct the annotations, the required domain knowledge to perform a given annotation task depends mainly on the underlying problem statement and the used imaging modality. Furthermore, the scope and form of high-quality labeling instructions may vary depending on the underlying problem. For simpler problems, such as eye color classification, shorter labeling instructions may be sufficient. Complicated annotation tasks, on the other hand, may require elaborate labeling instructions up to 400 pages in length [174] to properly communicate the necessary information. Regarding form, a labeling instruction is not restricted to being a document, and could be presented in video or app format as well. Which format or combination works best for the problem at hand should be determined individually. Regardless, the chosen format should be archivable to enable consistent onboarding of additional annotators and access for future users of the dataset to understand the data generation process. In case of doubt, sharing more information than necessary in the labeling instructions is preferable to sharing too little information, as evidenced by our work (see Figure 4.4).

Of note, with performance assessment of crowdworkers still conducted by examining their performance on reference standard datasets [51, 98] or measuring the inter-worker agreement [102, 136], the recommended iterative labeling instruction generation would potentially reduce errors resulting from lacking quality in labeling instructions and should thus result in a more precise assessment of crowdworkers' actual performances.

Another implication of our findings is the effect of annotation errors on the life cycle of a ML system. Data critically impacts the entire pipeline of a ML system, from the initial problem statement up to the final model deployment [168]. Annotation errors, often referred to as "label

noise", thus represent a long-standing challenge in the ML community, which has led to the development of preprocessing methods to clean data or the creation of models that are more robust to annotation errors [88]. This holds especially true for the safety-critical biomedical domain, in which high-quality test images are the foundation for the medical certification of new solutions [118, 161]. It should be noted that robustness also plays a particularly important role in this domain. Hence, there is a risk of rare cases being mislabeled and not contributing sufficiently to the robustness assessment. Our experiments show that images with rare but relevant characteristics particularly benefit from labeling instructions with a higher density of information (cf. Figure 4.5).

Furthermore, annotation errors in the test set have drastic consequences for ML competitions, which are often considered the gold standard for identifying the best algorithm for a specific research question. Competitions typically lead to winning algorithms becoming the new SOTA method and being awarded tremendous monetary rewards and recognition [114]. Even for well established datasets, such as ImageNet [94], annotation errors in the test set falsify the selection of the best performing model. For example, a trained ResNet-18 model was shown to outperform a trained ResNet-50 model if the prevalence of originally mislabeled test examples increases by just 6%, given the corrected ImageNet test set [135]. This issue is further aggravated by the fact that competition test sets are frequently inaccessible to the public after a competition ends, making auditing of the test sets impossible. A comprehensive overview of data issues and their cascading effect is provided by Sambasivan et al. [168]. In summary, reducing annotation errors early on within the life cycle of an ML system by providing higher quality labeling instructions positively impacts data preprocessing, model selection, model training, model validation and finally model deployment.

A limitation of our study could be seen in the fact that we only included one dataset. Our chosen dataset combines the advantages of high-quality reference annotations, representing the real-world complexity with gradually increasing stages of difficulty and high volume. We chose to only include one sample scenario since running several experiments with the same annotation companies and different datasets poses significant risks of exposing the experimental setting. Annotation companies are typically well aware of the most common datasets and would additionally spot the known experimental structure of gradually increasing labeling instructions within the same layout. Awareness of the experimental setting would in turn lead to results being skewed in favor of high-quality performances, since these companies are inherently motivated to present their work as reliable. A further limitation of our study is that MTurk required a different annotation tooling than the annotation tooling used by the professional annotation companies. To mitigate a potential impact of the toolings, all participating annotators had no prior experience with their respective tooling and both toolings included best design practices to enable high quality annotations.

Our study was subject to several design choices. For the performance measurement, we focused on the DSC as an overlap-based metric. Utilizing distance-based metrics, such as the Normalized Surface Distance, yielded similar results. For the statistical analysis, we assumed that the non-zero DSC scores follow a beta distribution as it is the more natural distribution for this kind of data. We further assumed that the random effects in the two-part ZIBMM are normally distributed and correlated. The correlation assumption was reasonable in that, as the probability of severe annotation errors increases, the expected DSC score decreases and vice versa.

Even though this paper focuses on labeling instructions associated with biomedical image analysis, we believe that the findings can be translated to other research fields, and to crowdsourced data annotation in general. We expect the impact to go beyond academia because industrial production ML projects by nature depend on a predefined level of annotation quality in order to obtain the required algorithm performance level. Consequently, it stands to reason that more effort and monetary resources should especially be invested in developing labeling instructions in industry [51].

The present work opens up several future research avenues: First, a structured competition submission system has improved the quality of biomedical image analysis competitions in recent years [115]. By collecting the submitted labeling instructions and the corresponding feedback over time, a model could be trained to provide an automated quality feedback mechanism for labeling instructions in the biomedical domain. Second, professional annotation companies usually conduct quality assurance checks by experienced annotators or team leads before providing their created annotations to the annotation requester. Although we obtained significant quality improvements in the obtained annotations by optimizing the labeling instructions, it would be of interest to analyze potential further impact of such quality assurance checks. Additionally, the interaction effects between experienced annotators working on quality assurance checks and instructions with varying levels of information density could be of interest to the scientific community. Third, data pipelines with a long-term focus face the risk of concept drift, where the initially captured distribution of input data changes. How labeling instructions should evolve along ever-changing data and its distribution remains an open question to be tackled. Finally, with an increasing educational shift towards digitization and data science, involving medical students in medical image annotation as part of their study program could potentially become a new source of crowdsourcing. Future research should examine this promising symbiotic relationship, where medical students obtain hands-on ML-related skills highly relevant to their profession while at the same time easing the annotation bottleneck for the scientific community.

In summary, our study is the first to examine the impact of the quality of labeling instructions on annotation work performed both by professional annotators and crowdworkers. We uncovered a significant discrepancy between the demand of professional annotators for better labeling instructions and current common practice. Given the rapidly increasing complexity and diversity of datasets, we envision the establishment and widespread adoption of quality standards for labeling instructions to become imperative in the future.

4.2 Quantifying the Impact of Quality Assurance

Author's Note on Reused Material:
This section and the referenced appendix sections substantially build upon the peer-reviewed publication: Tim Rädsch et al., "Quality Assured: Rethinking Annotation Strategies in Imaging AI", published in the *Proceedings of the European Conference on Computer Vision (ECCV)* 2024 [157]. As first and lead author, I conceived the study, coordinated all interactions with the four commercial annotation companies, specified and supervised the large-scale instance-segmentation campaign, and lead the implementation of the full experimental and statistical pipeline, including code for data pre-processing, quality-metric computation, and mixed-effects modeling. I designed the figures, secured the necessary annotation funding, and directed the interpretation of results. The manuscript was written primarily by me, with co-authors contributing domain-specific feedback, supplemental analyses, and iterative editorial revisions. Significant portions of the original text, tables, figures, and results have been adapted or expanded for integration into this chapter in order to maintain coherence with the surrounding thesis. Permissions for text and figure reuse have been confirmed with the publisher. The concept passage has been rewritten so that each experiment's rationale, finding, and implication is clearly explained, and the respective Discussion subsection enables readers to understand both the integrated contribution and its individual ramifications. This work was co-authored with Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Heller, Fabian Isensee, Annette Kopp-Schneider, and Lena Maier-Hein.

4.2.1 Concept Overview

Reliable benchmarking of image-analysis algorithms depends not only on clear, example-driven labeling instructions, as established in the previous chapter, but also on the *execution context* in which those instructions are applied. Having shown that instruction design is decisive for annotation accuracy, we now investigate two further determinants that arise once the content of the instructions is fixed: the *choice of annotation provider* and the *effectiveness of providers' internal quality-assurance (QA) pipelines*. This progression from *what* to instruct toward *where* and *how* to annotate constitutes the logical next step in constructing dependable reference datasets.

Our study encompasses 57 648 instance-segmented images annotated by 924 workers and reviewed by 34 QA specialists across four professional annotation companies and MTurk, as depicted in Figure 4.7. Annotation quality is quantified with the DSC and the Normalized Surface Distance (NSD), while intentional low-effort submissions are filtered through duration thresholds and expert screening. A logistic mixed-effects model, incorporating random effects for image, annotator, and provider, captures variance in QA efficacy across 57 636 metadata-labeled images.

We address three research questions:

- **RQ 1** - *Provider performance.* How does the choice of annotation provider influence the quality and cost of annotations?
- **RQ 2** - *Added value of internal QA.* Does annotation companies' quality assurance improve annotation quality?
- **RQ 3** - *Selective QA.* Should quality assurance focus on specific subsets of images?

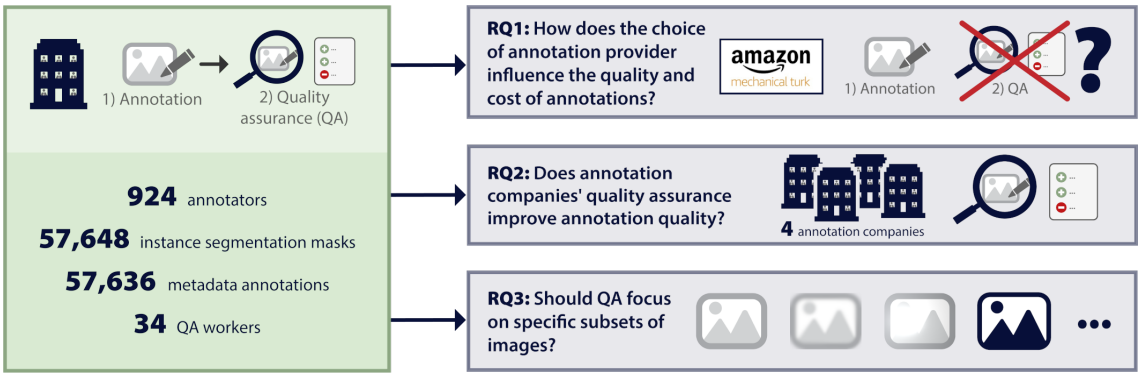


Figure 4.7: **Research Questions (RQs) tackled in this work.** Based on 57,648 instance segmentation masks annotated by 924 annotators and 34 quality assurance (QA) workers from five different annotation providers, we (1) compared the effectiveness of generating high-quality annotations between annotation companies and Amazon Mechanical Turk and (2) investigated the effects of annotation companies' internal QA and (3) real-world image characteristics on the annotation quality.

The results reveal that (i) annotation companies deliver higher annotation volumes and superior DSC/NSDscores than MTurk at lower cost; (ii) internal QA yields only marginal improvements once detailed, example-rich instructions are in place, whereas further investment in instruction engineering produces substantially larger gains; and (iii) QA benefits concentrate on visually challenging images, such as those with low illumination or complex object overlap, indicating that targeted rather than blanket QA maximizes cost-benefit efficiency.

Taken together, this work complements the previous chapter by showing that, after optimizing labeling instructions, the largest additional improvements in annotation quality arise from discreet annotation provider selection and selective, characteristic-aware QA. These insights complete a coherent framework for producing high-quality reference data and thereby reinforce the empirical foundations for trustworthy benchmarking and real-world deployment of image-analysis algorithms.

4.2.2 Methods

In the following, we describe the selected data, the labeling instructions, annotation providers and annotators, and the QA process of our work.

Definitions

We follow the definitions as in the previous contribution, see Section 4.1.2. This ensures consistency for our experiments and should ease the statistical analysis.

Analysis of Computer Vision Conference Exhibitors

We included the major computer vision conferences from 2022 and 2023, namely: ECCV 2022, CVPR 2022, ICCV 2023, and CVPR 2023. The list of exhibitors was extracted from the respective

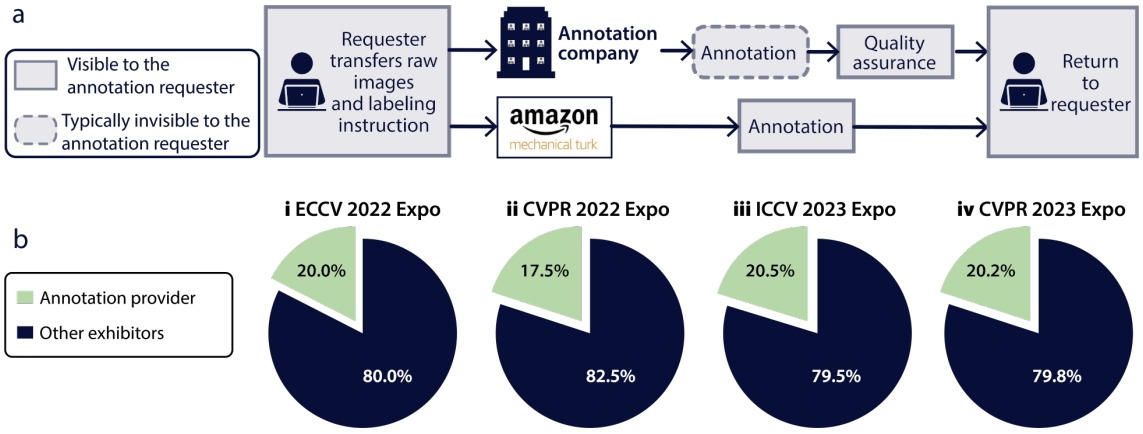


Figure 4.8: **a)** In contrast to Amazon Mechanical Turk (MTurk), annotation companies commonly perform quality assurance (QA) before delivering the result to the annotation requester. **b) Annotation providers are an integral part of the computer vision community.** They account for around 20% of exhibitors present at major computer vision conferences (European Conference on Computer Vision (ECCV) 2022: n = 40, Conference on Computer Vision and Pattern Recognition (CVPR) 2022: n = 97, European Conference on Computer Vision (ICCV) 2023: n = 43, CVPR 2023: n = 104) in the last two years.

conference websites[1]. We define an annotation provider as an entity that provides annotations performed by human workers. Consequently, each exhibitor's website was analyzed to determine whether the exhibitor provides human-generated annotations as a core component of their offering. Companies offering solely synthetic data were not included.

Data Selection

We conducted our experiments on the instance segmentation data from the HeiCo for surgical data science in the sensor operating room [116], representing a safety-critical application that crowdsourcing platforms can still handle. The task involved segmenting surgical instruments from laparoscopic videos, resulting in 57,648 instance segmentation masks on 4,050 unique frames. Furthermore, we used 57,636 metadata annotations on the unique images to quantify the effect of varying real-world image characteristics (see RQ3). Characteristics included motion blur, underexposure, object occlusion, or overlapping objects, among others, and are illustrated in Appendix A.8. Overall, the goal was to have a broad range of real-world image characteristics represented in our selected data.

Labeling Instructions

To investigate the benefit of QA compared to an investment in detailed labeling instructions, we created three distinct types of labeling instructions with varying levels of information density, namely, (1) minimal text, (2) extended text, and (3) extended text including exemplary pictures.

[1]Accessed 2024-03-04: https://hallerickson.ungerboeck.com/prod/app85.cshtml?aat=
{wRAm7m6yFc6aFoOwv9ggPME6PgM3DwEgvwcFfgyN3YY%3d;
Cr9Jl1poqL33RLclLrfrlBKRmhROPXGkDwcQAVv7zHM%3D;
Zva%2bsPD7WeTWWGsGWvarzixed2IJHWExCd84dKNS6Yg%3d};
https://iccv2023.thecvf.com/exhibitor.list.and.floor.plan-365000-5-44-50.php

Each subsequent labeling instruction type was structured to contain more detailed information than the preceding. To ensure accurate labeling of the imaging data, we established a collection of design principles applicable to all labeling instructions:

(1) Labeling information is presented in a slide format to enhance the processing of information by humans.
(2) Each slide contains a distinct 'chunk' of information, which is a self-contained piece of content. This approach of 'chunking' information lessens the cognitive load on the working memory.
(3) Related information chunks are placed in close proximity to one another.
(4) A uniform design template is used throughout, featuring specific fonts, symbols, and color schemes.

The three labeling instructions differ as follows:

Minimal text labeling instructions: These instructions provide a concise textual outline, offering examples of typical annotations and highlighting the most frequent annotation types (see Appendix A.5). They reflect a scenario with minimal effort invested in developing the instructions. The basic framework includes seven slides totaling 168 words.

Extended text labeling instructions: Building upon the simplified version, the extended text instructions offer a thorough textual explanation, supplemented by both positive and negative examples (refer to Appendix A.6). They cover both regular and some rare annotation scenarios. The enhanced version comprises ten slides with a total of 446 words.

Extended text including pictures labeling instructions: These instructions augment the enhanced textual guidelines with images (detailed in Appendix A.7). The visual aids feature descriptions, symbols, annotations, and color coding to effectively communicate the details on the slides. They also include illustrations of rare annotation cases, indicating a deep and well-documented understanding of the labeling process. This set contains 16 slides with a total of 961 words. We refer the reader to the Appendix for full details of each labeling instruction type. From a 2025 perspective, an increasing number of projects uses labeling instructions. We postulate that the demand for more complex data (see Section 3.1.2) is a major contributing factor. Overall, with increasing annotation complexity, the agreement between annotators falls, because more interpretation variants for the same data point are possible.

Annotation Provider Selection

Annotation provider selection should include a broad range of dimensions to accurately meet the requester's demand. In the following, we will highlight some relevant dimensions that influence annotation quality:
Price - Given a fixed annotation budget, the price per annotation determines the final annotation volume. However, price alone does not necessarily reflect annotation quality. Our research has shown that higher quality can sometimes come at a lower cost. Given the international distribution of annotation work, prices can vary widely for the same project. Therefore, price should be considered alongside other dimensions to ensure a comprehensive evaluation.

Annotator expertise - Overall, higher annotator expertise results in higher-quality annotations. This includes both (1) the necessary knowledge to perform the annotation task and (2) the handling of the annotation tool. It is important to note that expert annotators may not produce high-quality annotations if they lack tool-handling skills or sufficient time. Conversely, adequate

knowledge of the task is essential for generating high-quality annotations. Splitting a complex annotation task into simpler subtasks can be an effective strategy to reduce the knowledge required from each annotator.

Risk diversification - Large-scale annotation projects often span several months or even years. To mitigate risks, it is essential to diversify the annotation pipeline. Relying on a single country or provider can be risky, as demonstrated by various recent international crises. Diversifying providers across different regions can enhance the resilience of the annotation pipeline.

Ethics - Providing adequate working conditions for annotators should be a standard practice. We advocate for fair treatment and proper compensation for annotators, as demonstrated by our experiments. For a more in-depth discussion on the ethics of data annotation, we refer readers to the work of Miceli et al [125].

Annotation providers

The study was conducted based on 216 annotators and 34 QA workers from four annotation companies and 708 crowdworkers from the crowdsourcing platform MTurk. Our goal was to provide the best representation for both annotation provider types. To ensure comparability, we had selected the same providers as in the previous study. It is generally uncommon to employ professional companies for small scale projects that are limited to QA. However an increasing number of companies are founded that focus exclusively on QA [108]. For most scenarios, they are employed for larger projects with seven digit dollar project volumes.

Consequently, we established higher standards for our MTurk crowdworkers compared to the usual research benchmarks, which generally require a 95% acceptance rate of HITs and at least 100 completed tasks [25, 67, 15]. The goal was to generate the highest quality annotations on the one hand and keep them as close to a real world annotation project on the other hand. Therefore, our criteria included a 98% acceptance rate with over 5,000 completed HITs, surpassing the vast majority of all crowdsourcing studies in the literature. We spread out the HITs over 40 days at all times to obtain a representative MTurk sample and compensated workers fairly, as recommended by Litman et al. [111]. Despite its substantial impact on annotator performance, worker compensation is often overlooked by the community [125]. The annotation companies we selected have a solid background in handling large-scale annotation tasks in a variety of domains and operate internationally. The annotators were allocated by the companies themselves to replicate the typical workflow of a project carried out by an annotation company. Each annotator was linked to a specific annotation provider that had been predetermined. No selection process from our side took place for either the annotators, QA workers, or the MTurk crowdworkers, to prevent annotation companies to game our experiments.

To prevent information leakage, we conducted the labeling process with increasingly detailed labeling instructions (minimal text instruction, extended text instruction, and extended text including pictures instruction). Each annotator worked exclusively with one of the three instruction types as introduced in Section 4.2.2. To enhance security, we included a mandatory minimum gap of ten days between consecutive labeling instructions.

Annotation Tool

MTurk operates on a self-service basis. In this model, the individual or organization requesting annotations is responsible for supplying all aspects, including training for annotators and

the annotation tools, whereas the crowdworkers are engaged as independent contractors. In contrast, annotation companies provide a managed service for handling annotators, who can work on any annotation tool that the requester gives access to. Since the tool used for the annotation companies was not available for MTurk, we used a different annotation tool for MTurk. Both annotation tools were developed with best design practices to ensure the production of high-quality annotations. Furthermore, the tools were tested internally, until a sufficient user experience and annotation result were achieved. As intermediate annotations without QA are typically not visible to the requester, the software company added automated backend exports after the annotation step to enable this study. This enabled our study in the first place and decreased the chance for the annotation companies detecting our experiments. We even excluded the teamleads, which normally have a broader access in the project management software in comparison to regular annotators or QA workers. Professional annotators are aware that their work is internally verified by QA workers. None of the annotators participating had previous experience with the annotation tools they were using. This was the only possible solution, because companies use different tools on a regular basis. Including a more known tool, would give certain annotation companies an unfair advantage and increase the modeling difficulty.

Quality Assurance Process

In the internal QA process, annotators typically report to a small number of experienced QA workers, who are often responsible for training their team and, in our case, responsible for the QA of the annotations.

The general workflow for all companies consisted of an annotation stage, in which images were assigned randomly to each annotator, and a QA stage. Similarly, each of the annotators was assigned randomly to a QA worker. This resulted in around seven annotators per QA worker. During the annotation stage, annotators received raw images with the task of annotating them until they considered the quality of annotations satisfactory. Annotators were aware that their work would be checked by QA. However, they did not know in advance, which QA worker would check their work. Once submitted, the QA worker gained access to the generated annotations, allowing them to make modifications until they were satisfied before submitting the final annotations for delivery to the requester. Throughout this process, both annotators and QA workers retained access to the respective labeling instructions at any given moment. We obtained access to the annotations both before and after QA through a software company handling a substantial number of annotation projects annually. The experiments were seamlessly integrated into the software company's pipeline of annotation projects, reducing the risk of exposing our experiments. Pricing was established to be representative of other projects of similar size to risk the exposure potential and enable fair worker compensation.

To maintain fairness and avoid giving any annotation provider a potential informational advantage that may have affected our statistical analysis, no specific queries from the annotation companies about the content of the labeling instructions were addressed. While this might seem unintuitive at first, it is quite common, as evidenced by the small number of projects that even provide labeling instructions (see Figure 4.2).

4.2.3 Experiments

In the remainder of this section we present the experimental design conducted for each RQ. Study design, evaluation metrics, and statistical procedures are delineated individually for **RQ 1**-**RQ 3** as outlined below.

RQ 1: Annotation Provider Selection

Each of the five labeling providers annotated the same images with each set of labeling instructions, employing separate annotators for each task. We analyzed the number of spam images, images with severe errors, i.e., images with at least one false positive and/or false negative, and the overall annotation quality with the DSC [43] as the overlap-based metric and the NSD [161] as the distance-based metric. We define a spam annotation as a purposefully poor-quality annotation with the intent to obtain the payment without properly conducting the annotation task. Annotation duration was used as an auxiliary measure, where accessible. For instance, if an annotator completed the task in only 17 seconds, they could not have read the instructions and properly performed annotations on several images. Of note, a spammer could wait a relevant number of seconds and then enter a poor-quality annotation to bypass the annotation duration check. Thus, each image was additionally visually inspected by an engineer with extensive labeling experience. Spam annotation examples are displayed in Appendix A.9. We did not use other measures, such as mouse movement or other telemetry data, because they were not available via MTurk.

RQ2: Internal QA Analysis

The annotation companies conducted their internal QA processes directly after their annotators completed the annotation job, as explained in Section 4.2.2 and Figure 4.8a. The experiments began with only providing the minimal text labeling instructions, followed by the extended text labeling instructions, and finally, the extended text instructions with pictures, to prevent information leakage. A subsequent round of annotations was only

For each annotated image, we calculated the DSC and NSD per instance, aggregated the scores per annotated image, and aggregated the scores over each annotation company and each labeling instruction. We then compared the annotation quality without and with QA across the three types of labeling instructions for the four annotation companies by computing both absolute and relative changes in DSC/NSD aggregates, the percentage of modified images, and the change in the percentage of severe errors, among other parameters.

RQ3: Image Characteristics

In contrast to MTurk, the four annotation companies conducted an additional annotation round, including QA, with the extended text including pictures labeling instructions to quantify the impact of different real-world image characteristics, represented by 57,636 metadata annotations, on the probability of improvement. The data was analyzed with a logistic mixed model [123] which included random effects for the image, the annotator, and the annotation company and the model was implemented in the lme4 package in R. The obtained estimates of the covariates are on the log-odds scale and were exponentiated to obtain the odds ratio for each covariate. The implementation was conducted with the programming language R (Software: R version

4.0.2, package lme4 version 1.1.33). Similar to the previous contribution, we included statisticians before the start of the project to enable a representative data selection, overall annotation and QA setup and resulting statistical analysis.

$$\text{I} \sim (1|\text{Image\_name}) + (1|\text{Company/Annotator}) + \sum_{i=1}^{7} G_i + \sum_{i=1}^{10} D_i \tag{4.4}$$

where:

- Improvement (I) is the response variable.
- $(1|\text{Image\_name})$ represents a random intercept for each Image_name.
- $(1|\text{Company/Annotator})$ represents a random intercept, with Annotator nested within Company.
- $\sum_{i=1}^{7} G_i$ represents the generalizable fixed effects, with G_i being the i^{th} analyzed generalizable covariate.
- $\sum_{i=1}^{10} D_i$ represents the domain-specific fixed effects, with D_i being the i^{th} analyzed domain-specific covariate.

List of covariates:

- Generalizable covariates:
  - Motion artifacts in the background
  - Other objects present in the background
  - Object occluded by background
  - Intersection of objects
  - Motion artifacts on the object
  - Object(s) overexposure
  - Object(s) underexposure
- Domain-specific covariates:
  - Background covered by blood
  - Background covered by reflections
  - Background covered by smoke
  - Image overlaid with text
  - Caption at the bottom of the image
  - Black box present in the image
  - Trocar present in the image
  - Object covered by blood
  - Object(s) covered by reflections

- Object(s) covered by smoke

Real-world image condition analysis
For the impact analysis of real-world image characteristics on QA, we focused solely on the extended text including pictures labeling instructions and the annotation companies. While it would be interesting to observe the interaction with the other labeling instructions as well, the companies were already exposed to the instruction with the highest information density. Therefore, they would have access to additional information that would prevent a fair comparison between the different instructions. All 4,050 unique frames of the HeiCo dataset were annotated two times by each annotation company, accumulating eight annotations for each frame. For this analysis we utilized all collected 57,636 metadata annotations.

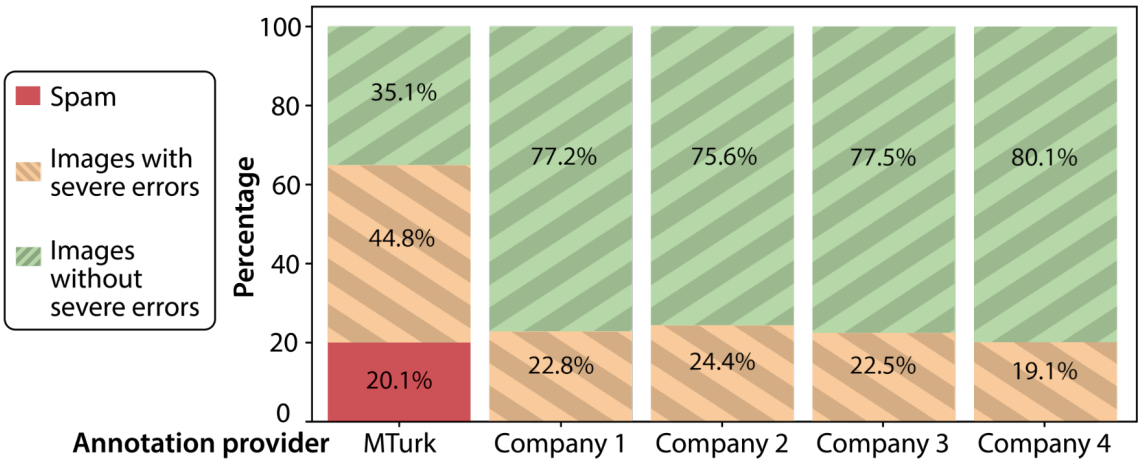


Figure 4.9: **Annotation companies are more efficient at generating high-quality images than Amazon Mechanical Turk (MTurk).** Percentage of images without severe errors (green), images with severe errors (orange), and spam (red) obtained for the five different annotation providers investigated. Annotation costs per image were lower for all companies compared to MTurk (median: 61.1%).

4.2.4 Results

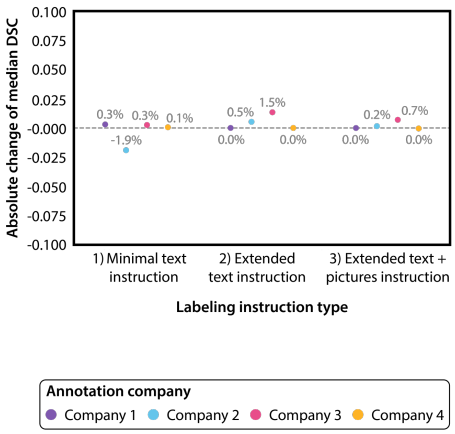
Annotation Providers Are a Crucial Part of The Computer Vision Community

Based our computer vision conference exhibitor analysis, we were able the identify their impact for the scientific community. In the last two years, around 20% of the conference exhibitors were annotation providers, as illustrated Figure 4.8. As the majority of exhibitors were offering specific tech products, we excluded exhibitor booths from our analysis that were not clearly related to a specific offering (e.g., the booth for the ECCV coffee station or the Springer Publishing Group). This resulted in three excluded exhibitors for ECCV 2022 (total before exclusion: 43), seven excluded exhibitors for CVPR 2022 (total before exclusion: 104), three excluded exhibitors for ICCV 2023 (total before exclusion: 47), and eight excluded exhibitors for CVPR 2023 (total before exclusion: 112). At the time of the study, computer vision research was considered one of the most in-demand research topics, which might explain the large number of vendors at CVPR in contrast to the other major ML conferences. Since then, interest in LLMs and VLMs has risen rapidly with the release of ChatGPT in 2023. This might imply that vendors are shifting to other conferences that go beyond computer vision.

Professional Annotation Companies Outperform MTurk Both in Terms of Cost and Quality of Annotations

Annotated images generated by the annotation companies were of higher quality, comprising no spam (MTurk: $\sim$20%) and a much higher proportion of remaining images without severe annotation errors. In particular, we observed a factor of > 2 improvement for all companies compared to MTurk, as shown in Figure 4.9. In addition, the observed annotation companies were 61.1% cheaper in the median, despite their additional employment of QA workers. More specifically, the costs ranged from 24.5% to 81.5% of the MTurk costs for the same images and number of annotations. We have not been given permission to reveal the identities and exact

(a)



(b)

Company	Labeling instruction	25th Percentile Annotation	25th Percentile QA	Median Annotation	Median QA	75th Percentile Annotation	75th Percentile QA
1	1) Min. text	0.61	0.62	0.93	0.93	0.96	0.96
	2) Ext. text	0.58	0.58	0.93	0.93	0.96	0.96
	3) Ext. text + pictures	0.90	0.90	0.95	0.95	0.96	0.96
2	1) Min. text	**0.57**	**0.47**	0.92	0.90	0.96	0.95
	2) Ext. text	**0.62**	**0.84**	0.93	0.94	0.96	0.96
	3) Ext. text + pictures	0.85	0.88	0.94	0.94	0.96	0.96
3	1) Min. text	0.60	0.60	0.93	0.93	0.96	0.96
	2) Ext. text	**0.47**	**0.64**	0.91	0.93	0.96	0.96
	3) Ext. text + pictures	**0.74**	**0.85**	0.93	0.94	0.96	0.96
4	1) Min. text	0.63	0.63	0.94	0.94	0.96	0.96
	2) Ext. text	0.88	0.88	0.94	0.94	0.96	0.96
	3) Ext. text + pictures	0.91	0.91	0.95	0.95	0.97	0.97

COLOR SCALE: 1.00 | 0.75 | 0.50 | 0.25 | 0.00 | -0.25 | -0.50 | -0.75 | -1.00

Figure 4.10: **QA of annotation companies only provides marginal improvements, if any. These depend highly on the company and type of labeling instruction. a:** Absolute change of the median DSC resulting from performing QA. For each company and labeling instruction, the median was obtained by aggregating the DSC scores over the images. **b:** Corresponding absolute values of the DSC. Positive and negative changes from annotation to QA are colored according to the color scale.

absolute budgets used for the annotation companies, hence we report costs relative to the MTurk budget. The companies consider their pricing strategies as a competitive advantage and include Non-disclosure agreements (NDAs) as part of their contracts. To simulate a regular annotation project, we agreed to these common practice. Examples of spam annotations are provided in Appendix A.9.

Regardless of the annotation scenario, the annotation companies generated higher quality annotations than MTurk with higher DSC scores including smaller IQRs (see Figure 4.14a), higher NSDscores including smaller IQRs (see Figure 4.14b), and lower numbers of severe errors (see Figure 4.14c). In fact, the annotation companies more than halved the number of images with severe annotation errors. While MTurk produced 44.8% images with severe errors, the best company generated less than 20%. Across all providers, the extended text including pictures instruction generated the lowest number of severe errors. While the extended text instruction did not improve the overall annotation quality, they still reduced the number of severe errors for three of the five observed providers.

Annotation Companies' Quality Assurance Does Not Improve Annotation Quality

As shown in Figure 4.10, the internal QA conducted by annotation companies does not substantially improve annotation quality if the QA workers only use minimal text instructions. The impact of QA when having access to extended text instructions or extended text including pictures instructions depended highly on the annotation company, while the effect was still minimal compared to not performing QA: Two of the four annotation companies show no change in the median DSC when comparing QA to non-QA outputs for extended text or extended text and picture instructions. The remaining two companies only show a very small relative improvement of below 2%.

We obtained even smaller improvements with the NSDscores, as depicted in Table 4.1. In certain

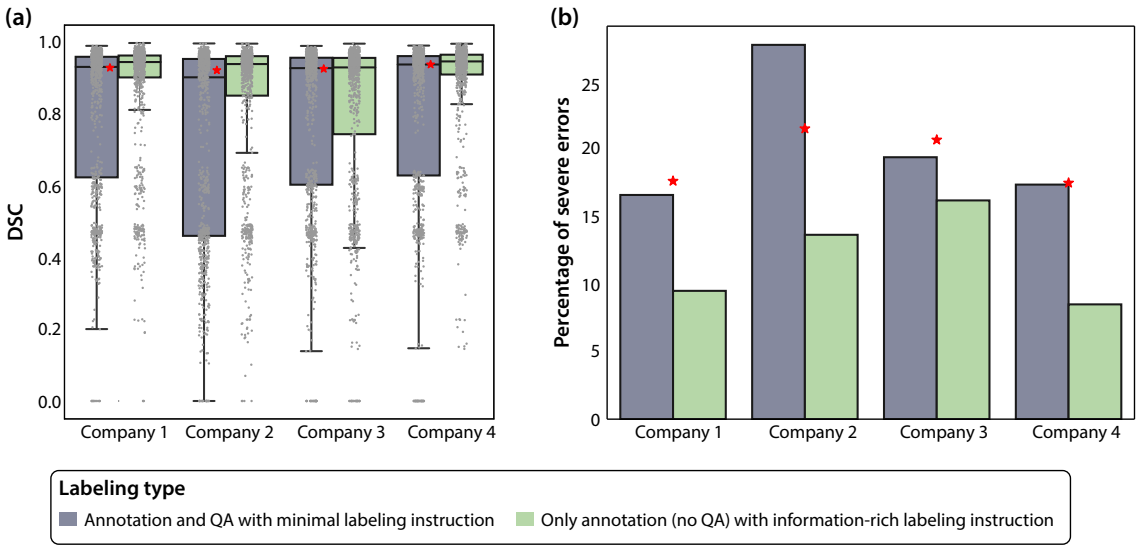


Figure 4.11: **Improving the labeling instructions yields a higher effect compared to adding QA to the annotation process.** The baseline performance (minimal labeling instruction, no QA) according to the DSC (a) and the percentage of severe errors (b) is shown individually per company with a red asterisk (minimal labeling instruction, no QA). While adding QA only yields marginal improvements (dark blue), major performance boosts are obtained by improving the labeling instructions provided to the annotators (green). DSC scores were aggregated for each image and per labeling type. DSC displayed as dots and box plot (band indicates the median, box indicates the first (25th percentile) and third (75th percentile) quartiles). $0 \leq \text{DSC} \leq 1$.

scenarios, QA was able to remove a small proportion of severe errors, however, the vast majority of severe remains in the data (see Table 4.2). Notably, the percentage of modified images by QA workers depended highly on the company, as depicted in Section 4.2.4. Company 4 modified less than 2%, yet they obtained the highest DSC score for the minimal text and extended text labeling instruction. For the third labeling instruction they modify the largest number of images. Figure 4.11 further illustrates that providing annotators with high-quality labeling instructions including pictures (green) leads them to vastly outperform annotators working with lower-quality labeling instructions, even if these are followed by additional internal QA (dark blue).

QA Workers Modify Only a Tiny Fraction of The Annotated Images

The number of modified images depended highly on the annotation company, as displayed in. Figure 4.12a. Nevertheless, a substantial proportion of modified images did not inherently imply superior annotation quality (Figure 4.12b). For instance, Company 1 on average made modifications to every third annotated image, while Company 4 only made modifications to every 10th image. However, it is evident that Company 4's resulting annotations clearly outperformed those of Company 1 when employing the extended text and extended text including pictures labeling instructions. The combination of a low modification rate and high annotation quality achieved by Company 4 could suggest a higher level of annotator expertise.

Overall, the annotations conducted with the extended text including pictures labeling instructions outperform all other experiments and QA efforts. This is a clear indication that further research

Table 4.1: **QA of annotation companies yields only marginal improvements in Normalized Surface Distance (NSD) scores.** The table reports NSDvalues at the 25th percentile, median, and 75th percentile for four companies with different labeling instruction (minimal text, extended text, extended text with pictures), comparing initial annotations and QA. Across all companies and instruction types, the differences between annotation and QA stages are small, indicating that QA provides limited additional benefit once detailed instructions are established. The values range from 0 (no agreement) to 1 (perfect agreement).

		25th Percentile		Median		75th Percentile	
Company	Stage	Annotate	QA	Annotate	QA	Annotate	QA
Company 1	1) Minimal text	0.65	0.66	0.99	0.99	1	1
Company 1	2) Extended text	0.62	0.61	0.99	0.99	1	1
Company 1	3) Extended text + pictures	0.97	0.97	0.99	0.99	1	1
Company 2	1) Minimal text	0.61	0.49	0.98	0.97	1	1
Company 2	2) Extended text	0.66	0.92	0.99	0.99	1	1
Company 2	3) Extended text + pictures	0.93	0.95	0.99	0.99	1	1
Company 3	1) Minimal text	0.63	0.65	0.99	0.99	1	1
Company 3	2) Extended text	0.5	0.67	0.98	0.99	1	1
Company 3	3) Extended text + pictures	0.82	0.93	0.99	0.99	1	1
Company 4	1) Minimal text	0.66	0.66	0.99	0.99	1	1
Company 4	2) Extended text	0.96	0.96	0.99	0.99	1	1
Company 4	3) Extended text + pictures	0.98	0.98	1	1	1	1

on internal QA processes is needed.

Quality Assurance Should Focus On Specific Subsets of Images

As shown in Figure 4.13a, QA significantly increases the odds of improving the annotation quality for images with generalizable image characteristics such as underexposed objects, intersection of objects, objects occluded by background in comparison to the absence of this characteristic (vertical black line). The only exceptions are motion artifacts present in the background and object(s) overexposure. The largest odd increase can be observed for intersecting objects, which is the only characteristic that includes more than one object of interest. This indicates that interacting objects represent one of the most difficult annotation cases. Furthermore, this characteristic can be added with a simple heuristic and does not need any human annotations.

In contrast to the general increase of images with generalizable image characteristics, none of the domain-specific image characteristics yield a significant change in the odds of improving the annotation quality over a regular image, see Figure 4.13b. The only exception for domain-specific characteristics is object(s) covered by smoke, which could be interpreted as a form of object occlusion.

4.2.5 Discussion

The data annotation market is estimated to reach a volume of more than USD 4.1 billion by the end of 2024 [31]. As the annotation process becomes increasingly automated and foundation models

Table 4.2: **QA of annotation companies yields only marginal reductions in severe error rates.** The table reports the percentage of severe errors for four companies with different labeling instruction (minimal text, extended text, extended text with pictures), comparing initial annotations and QA. Across all companies and instruction types, the reductions are small and inconsistent in direction, indicating that the effect of QA on severe errors depends strongly on both the company and the type of labeling instruction.

Company	Stage	Annotate	QA	Difference
Company 1	1) Minimal text	0.18	0.17	-0.01
Company 1	2) Extended text	0.2	0.2	0
Company 1	3) Extended text + pictures	0.1	0.1	0
Company 2	1) Minimal text	0.22	0.28	0.06
Company 2	2) Extended text	0.19	0.13	-0.06
Company 2	3) Extended text + pictures	0.14	0.11	-0.02
Company 3	1) Minimal text	0.21	0.19	-0.01
Company 3	2) Extended text	0.25	0.17	-0.08
Company 3	3) Extended text + pictures	0.16	0.12	-0.04
Company 4	1) Minimal text	0.18	0.17	0
Company 4	2) Extended text	0.12	0.12	0
Company 4	3) Extended text + pictures	0.09	0.09	0

evolve, the market is undergoing drastic shifts. With lower entry barriers for annotation, more efficient data annotation tools, and a rising impact of QA, it is thought to result in an estimated USD 8.2 billion market by 2028 [162]. As a research community, we need to understand how to ensure high annotation quality throughout this evolution, as reliable reference annotations are an indispensable foundation for the benchmarking of image analysis algorithms, and thus their ultimate translation into real-world practice.

To our knowledge, this study is the first to quantitatively and critically examine the impact of internal QA conducted by annotation companies in the context of safety-critical applications. The most important insights can be summarized as follows:

1. Annotation companies perform better both in terms of quantity and quality compared to the widely used platform MTurk (Figure 4.9). This implies that requesters working with crowdsourcing platforms should seriously consider annotation companies as an alternative. The organizational structure of these companies, which enables a single point of contact for requesters and eliminates spam annotations, appears to be a contributing factor to their efficiency. In light of poor working conditions being a major problem in annotation generation [125, 79], annotation companies, which can more easily be inspected than distributed crowdsourcing platforms, are thought to offer better working contracts, enabling fairer working conditions for annotators. It should be noted that not every annotation company automatically ensures adequate working conditions. The annotation companies employed in this study underwent thorough scrutiny to guarantee compliance with good working condition guidelines. With annotators working at a shared location instead of in a distributed crowd, verifying annotators' working conditions has become substantially easier. We thus view our work as an initial step towards greater account-

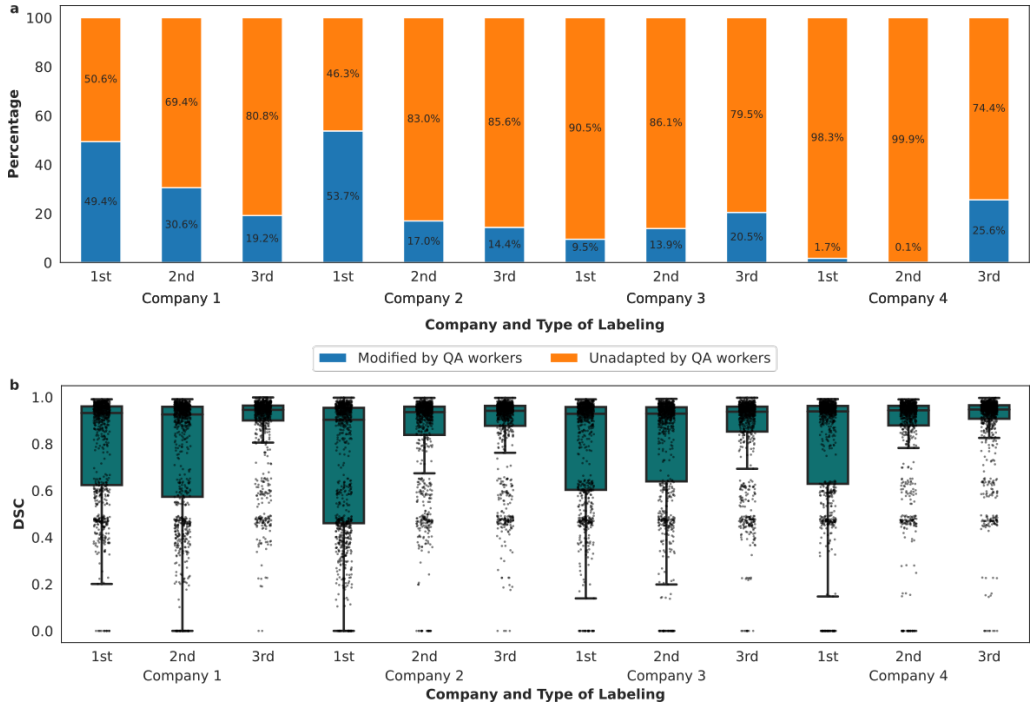


Figure 4.12: **The percentage of modifications by QA workers differed greatly between the companies. However, this did not necessarily result in higher annotation quality, when compared to other annotation companies. a**, The percentages of modified and unadapted images by QA workers. An image counts as modified if the responsible QA worker adapted the received instance segmentation mask from the annotator. **b**, The DSC was aggregated for each resulting annotated image and is displayed aggregated for each pair of annotation company and labeling instruction as a dots- and boxplot (the band indicates the median, the box indicates the first and third quartiles and the whiskers indicate $\pm 1.5\times$ interquartile range), the DSC maximum is 1 and the minimum is 0 for each image.

ability on the part of annotation requesters. Further dimensions of annotation provider selection are discussed in Section 4.2.2.

2. In contrast to common belief, internal QA mechanisms do not necessarily improve annotation quality (Figure 4.10). Furthermore, we observe no systematic QA quality improvement difference between annotation companies depending on their region of origin, size, or price, meaning more expensive annotation companies do not translate to better QA. However, improving labeling instructions instead of investing in QA can substantially boost the annotation performance (Figure 4.11). Consequently, in order to obtain high-quality annotations, annotation requesters should prioritize refining their labeling instructions, followed by implementing QA or even a different external QA process as a secondary step. Compared to QA, better labeling instructions can be generated with relatively few resources.

3. The benefit of QA depends on the specific image characteristics (Figure 4.13). In fact, QA did improve the quality of images with certain challenging real-world image characteristics, see Figure 4.13a. Consequently, it might be useful for requesters and annotation companies alike to allocate resources to identifying challenging image characteristics during a project, train

classifiers to identify images containing these characteristics and focus the internal QA effort on these images.

One could contend that the recent rise of foundation models marginalizes the role of annotation providers. We argue, however, that it only changes their role. In this context, it is necessary to differentiate between the data quality demands for the training and testing of models. With a recent trend towards fine-tuning pre-trained models, smaller high-quality datasets can achieve similar performance with less computation [101, 22], highlighting the importance of the quality of unlabeled data in unsupervised training and its role in reducing carbon emissions [186]. In testing, especially for safety-critical applications such as autonomous driving and clinical medicine, ensuring high-quality annotated test data is crucial for accurate model evaluation and real-world applicability [168]. Consequently, while training data annotation may lose relevance, test data annotation and its high quality demands are here to stay. Regarding foundation models, it should further be noted that annotations purely produced by foundation models introduce artifacts and biases [14] and represent an increasing rising problem for MTurk, as demonstrated by Veselovsky et al. [194].

Our study is subject to several limitations which deserve further discussion. Comparing prices across providers is non-trivial as the costs to annotate a given dataset depend on many factors including, among others, image quantity, type of annotations, location of annotation company and quality requirements. Our analysis was based on the actual costs incurred during our project. Confidentiality agreements with the software company which integrated our experiments into their annotation project pipeline prevent us from disclosing the identities of the annotation companies. Additionally, we took precautions to protect the identities of individual annotators and QA workers. This was particularly important given our findings that QA workers did not significantly enhance annotation quality, potentially exposing them to negative repercussions.

Another limitation could be seen in the fact that our dataset only represents one safety-critical domain. While this is a valid concern, we would like to highlight that we chose a dataset which combined quality, complexity, and volume. Furthermore, regarding the breadth of annotation providers and the number of annotators involved, our study is likely the largest of its kind, thereby incurring substantial costs.

Finally, conducting multiple experiments with identical annotation companies and varying datasets could have potentially revealed our experimental design. Had the annotation companies been aware of our research, outcomes could potentially have been biased to appear more favorable, given companies' inherent interest in demonstrating their services as reliable. We thus chose to conduct our experiments such that they could be seamlessly integrated into the software company's annotation pipeline, reducing their likelihood of exposure.

With the recent focus on VLMs, extending experimental analyses to multimodal data has the potential to generate novel insights of direct relevance to the research community. A complementary line of research concerns the use of expert-annotated data, which may provide a deeper understanding of the effectiveness of QA strategies on the current demand. We hypothesize that the observed effects will be of comparable magnitude, since expert annotators in niche domains often display substantial variability in their approaches and decision criteria. n additional question arises as to whether the presence of domain experts is strictly required for effective QA, or whether certain components of the process might be reliably delegated to non-expert annotators.

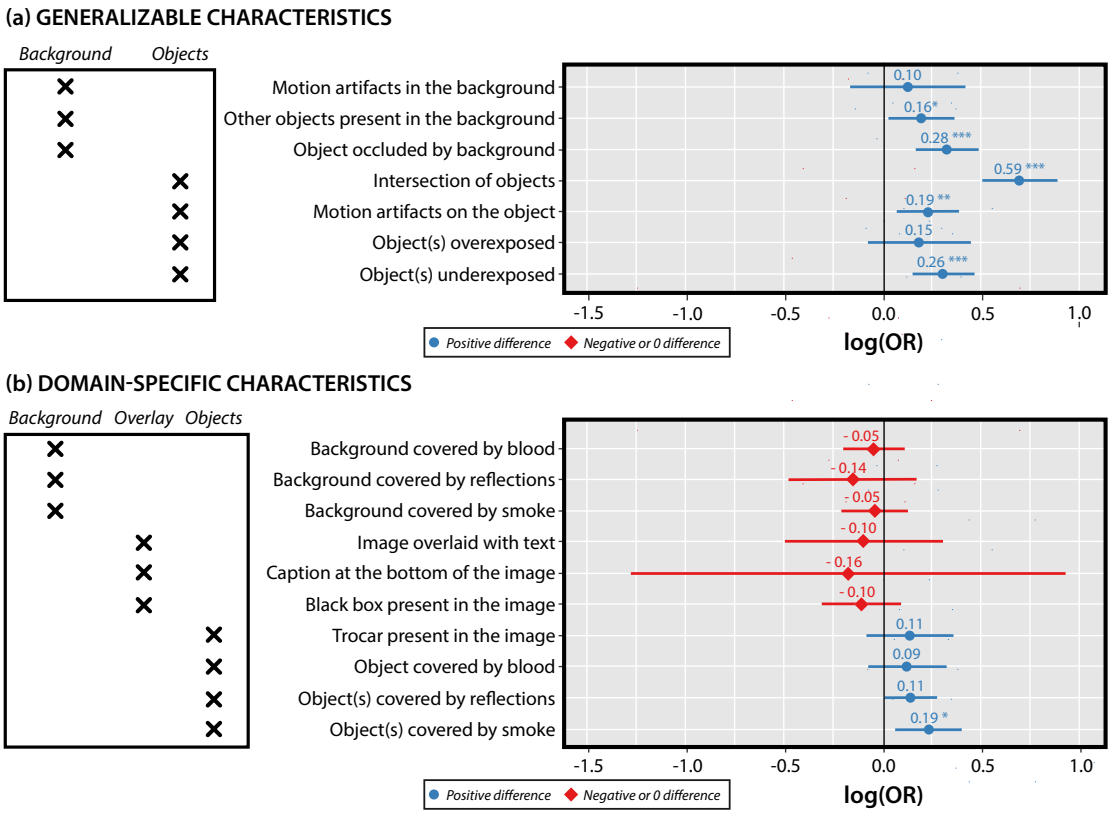
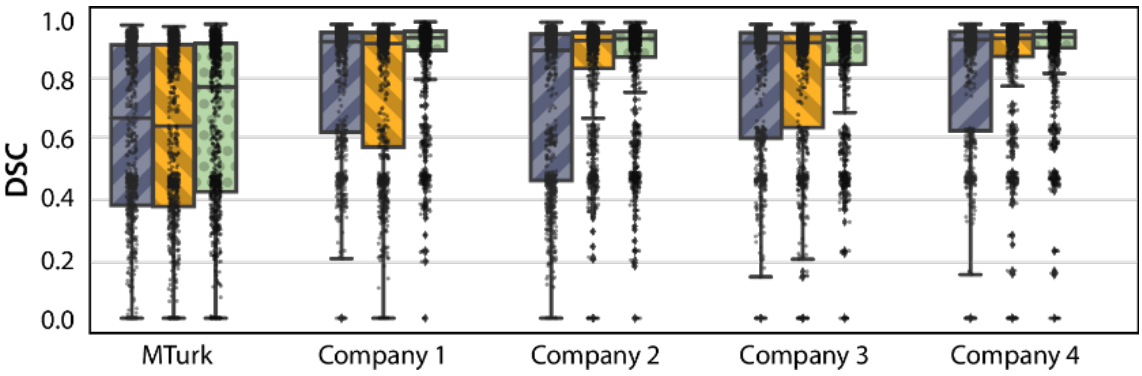
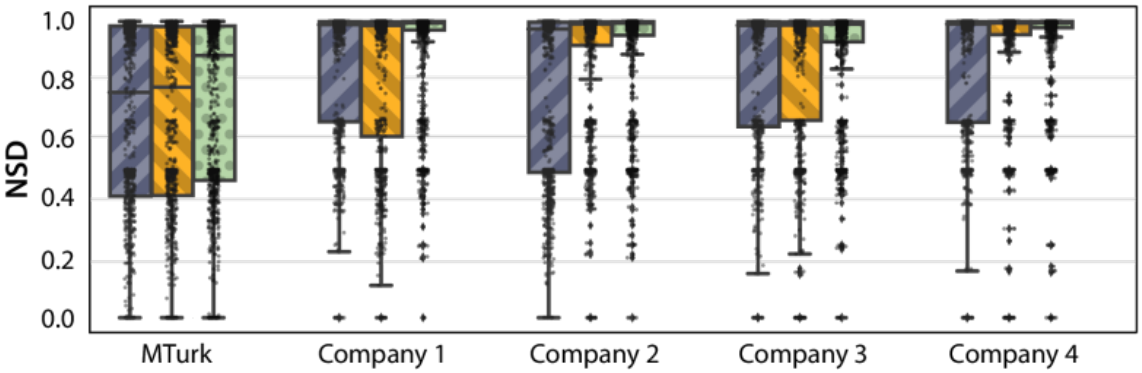


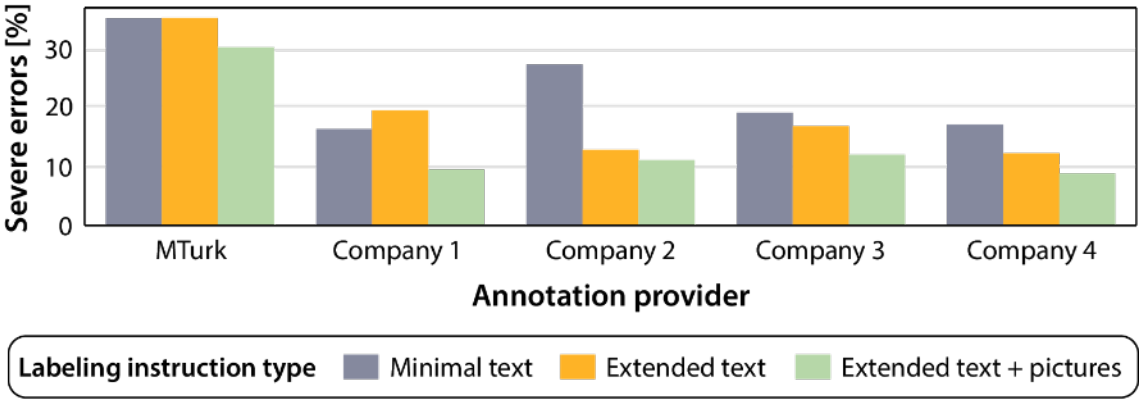
Figure 4.13: **QA of annotation companies can be beneficial to certain images with difficult real-world generalizable image characteristics, such as underexposure or intersection of objects.** We display the impact of (a) real-world generalizable and (b) domain-specific image characteristics on the odds of improvement. Improvement is defined as a positive difference in the QA, while no improvement refers to 0 or negative difference. The effect of an image characteristic is shown in the form of the log(odds ratio (OR)) representing the logarithms of the odds of improvement in the presence of the characteristic, compared to the odds of improvement occurring in the absence of that characteristic. The vertical black line refers to the point of equal odds. (Note: ***: p-value $\leq$ 0.001,**: p-value $\leq$ 0.01, *: p-value $\leq$ 0.05).



(a) Annotation companies generate higher DSC scores than MTurk. $0 \leq$ DSC ≤ 1.



(b) Annotation companies generate higher NSDscores than MTurk.$0 \leq$ NSD ≤ 1.



(c) Annotation companies generate a lower number of severe errors.

Figure 4.14: **Annotation companies generate higher quality annotations with their annotation pipeline than MTurk.** This is observed for (a) the DSC, (b) the NSD, and (c) number of severe errors (lower is better). Annotation companies include a QA step, in contrast to crowdsourcing on MTurk. (a) and (b) displayed as dots and box plots (band indicates the median, box indicates the first (25th percentile) and third (75th percentile) quartiles).

4.3 DomainBench: Generating VLM Benchmarks for Domain Researchers

This section and the referenced appendix sections are based in large parts on the submitted manuscript: Tim Rädsch et al., "DomainBench: A Framework for Scalable VLM Perception Benchmarking on Your Own Images", submitted to the*Conference on Neural Information Processing Systems* in 2025 (preprint: [156]). As the lead and first author of this publication, I played the central role in designing the study, preparing and conducting the annotation experiments, generating the benchmark datasets and implementing the associated code. I initiated and coordinated the collaboration with the involved QA company and annotators. I led the creation of the data enrichment pipeline and the benchmark pipeline. I led the interpretation of results and led the writing of the manuscript together with several co-authors, who provided analytical input, textual contributions, and iterative feedback. Significant portions of the text, figures, and results have been adapted and integrated into this chapter. Text and figures are extracted from the preprint at the time of submission. The concept passage has been rewritten so that each experiment's rationale, finding, and implication is clearly explained, and the respective Discussion subsection enables readers to understand both the integrated contribution and its individual ramifications. This work was co-authored with Leon Mayer, Simon Pavicic, A. Emre Kavur, Marcel Knopp, Barış Öztürk, Klaus Maier-Hein, Paul F. Jaeger, Fabian Isensee, Annika Reinke, and Lena Maier-Hein.

4.3.1 Concept Overview

Rigorous validation of VLMs constitutes a prerequisite for both scientific advancement and practical deployment. Benchmarking narrow, single task models faces numerous challenges [133], which grow substantially for generalist VLMs.

We observe three key trends that highlight the critical need for personalized benchmarking approaches:
Domain-specific benchmark demand:
Existing VLMs benchmarks provide valuable but predominantly general insights into model capabilities. Their heterogeneous design choices and the narrow emphasis on only a few imaging domains create substantial obstacles for systematic cross-domain comparison and for robust domain-specific assessment from a user-centric perspective.
Popular arena platforms do not scale from an individual user's perspective:
To gather a broad user representation, arena-style platforms, such as LM Arena [26], crowdsource single task submissions and answer rankings from users around the world. This setup supports personalized and domain-relevant validation, yet conducting large-scale assessments from a single-user perspective remains infeasible.
Homogeneous evaluation:
Most existing VLM benchmarks [52, 214, 201, 134] evaluate models using a single question per image. While this is common practice, it utilizes only a fraction of the image content, and fails to reflect domain users that typically have a small number of images and require evaluation across multiple aspects of each image.

To overcome these limitations, we introduce three contributions in this chapter: (1) DomainBench, a framework enabling resource-efficient construction of user-centric VLM benchmarks via task augmentation, which generates multiple diverse tasks from a single instance segmented image; (2) the release of seven new VLM datasets, developed according to a homogeneous protocol and

incorporating 142,333 carefully human-validated answers; and (3) a large-scale benchmarking study of 22 SOTA VLMs on 32,771 tasks, which exposes significant performance heterogeneity across domains and task categories, thereby substantiating the necessity of user-centric benchmarks. By adopting this methodology, researchers can achieve cost-efficient domain-specific model selection on their own images. The overview is displayed in Figure 4.15 and the DomainBench framework in Figure 4.16.

The results show that (i) open source models reduced the gap to closed source models, which remain in the leading position; (ii) models display high performance variability across domains for the same tasks; (iii) VLMs produce model family-specific capability weaknesses; (iv) human annotators generate errors predominantly on ambiguous or visually challenging questions.

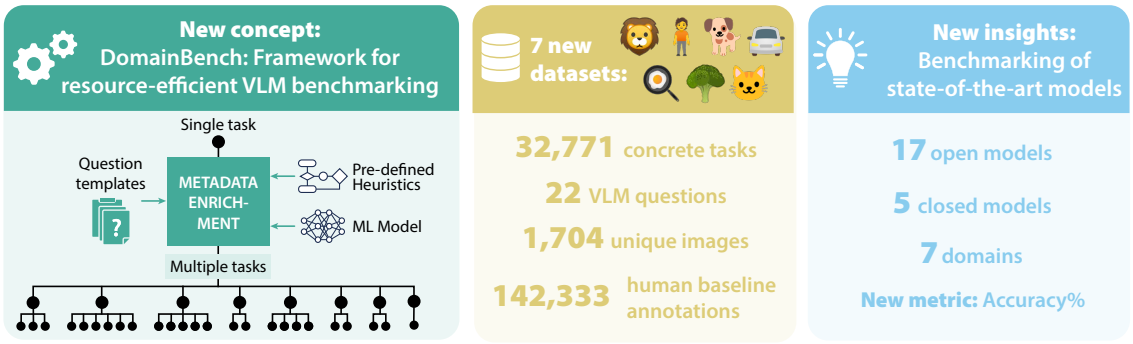


Figure 4.15: **Summary of contributions**. (1) New concept: We propose DomainBench, a new framework for the resource-efficient creation of domain-specific VLM benchmarks. It is based on the concept of task augmentation designed for creating multiple tasks from a single existing task using metadata annotations from automated sources, namely pre-defined heuristics applied to segmentations and existing robust generalist models. (2) 7 new datasets: We apply DomainBench to generate seven domain-specific VLM benchmarks with highly reliable reference data. As a unique feature compared to existing benchmarks, we quantify the ambiguity of each question for each image by acquiring human answers from a total of six raters. (3) New insights: We apply our framework to a total of 22 open and frontier closed models to demonstrate the benefit of task augmentation and to shed light on current VLM capabilities.

4.3.2 Methods

In the following, we describe the framework for resource-efficient in-domain benchmarking, the seven newly generated datasets, and the benchmarking strategy and further analysis details.

Conference Dataset Paper Analysis

To generate a comprehensive overview of the current dataset usage in the general computer vision field, we analyzed all papers from CVPR 2024 using three different LLMs. If the majority of models indicated that a paper introduced a new or modified dataset, we tagged it accordingly. This process identified 397 publications proposing a new or modified dataset. To validate the accuracy of the tagging, we randomly selected 10% of these flagged papers for a human review. All human-verified publications were confirmed to propose a new dataset. To exclude false negatives, we randomly selected an equal number of the unflagged papers and confirmed that they do not provide a new dataset.

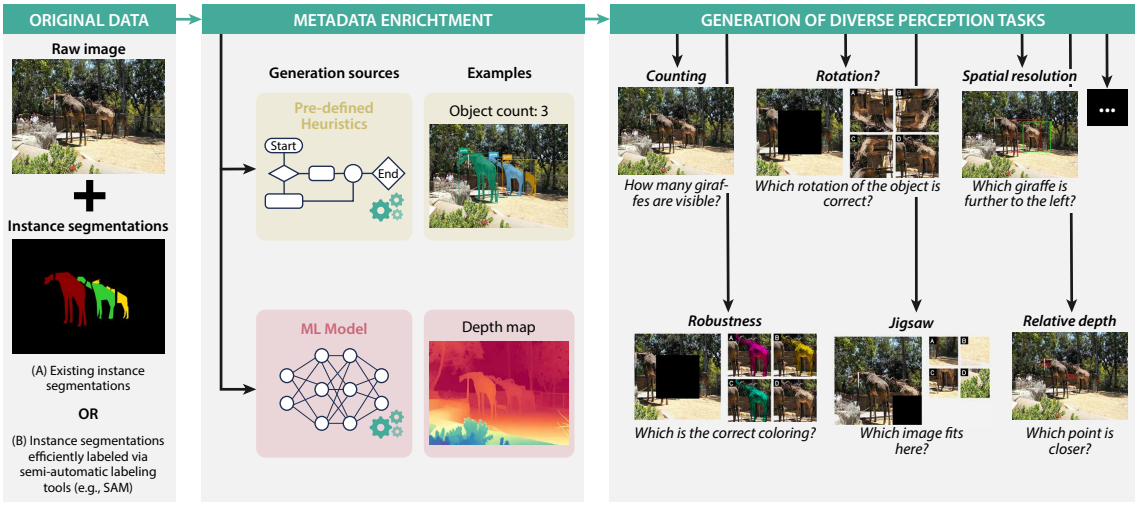


Figure 4.16: **DomainBench: A framework for resource-efficient in-domain benchmarking.** Starting from a single task with fine-grained annotations (here: instance segmentations), metadata for each image is obtained using pre-defined rules (e.g. inference of an object count) and information extracted with existing robust generalist models (e.g. relative depth information). This process transforms the initial task into a collection of 22 tasks, enabling resource-efficient and easy to use in-domain benchmarking of general VLM capabilities while maintaining cross-domain comparability.

DomainBench

The framework DomainBench for resource-efficient in-domain benchmarking is depicted in Figure 4.16. Starting with domain images that include instance segmentations (existing or created with semi-automatic labeling tools, such as SAM [91]), metadata for each image is acquired from automated sources, namely pre-defined heuristics applied to segmentations and existing robust generalist models to transform the single task into a collection of perception tasks.

For our seven new datasets, we use existing instance segmentations as the core perceptual task to generate the diverse set of VLM benchmark tasks depicted in Appendix B.2 (examples in Figure 4.22 and more detailed in Appendix C.2). Overall we aimed for a broad diversity of tasks on the object and image level that can be applied to any image with instance segmentation, regardless of the domain.

The metadata enrichment is derived from two sources:

1) Pre-defined heuristics and rules were employed to transform existing information into metadata. For example, instance segmentations were utilized to quantify the number of objects within a specific class or to determine whether specific instance segmentation masks were touching each other.

2) An existing depth foundation model, Depth Anything v2 [208], was used to generate depth maps for each image. This includes both the metric depth and the relative depth. Of note, future depth models can be integrated in our method seamlessly.

Existing Annotations	
Attribute	**Description**
relative_size	Relative size compared to image size
bbox_touches_bbox	Bounding box touching another bounding box
segmask_touches_segmask	Segmentation mask touching another segmentation mask
segmask_touches_segmask_with	Specific segmentation masks touching each other
segmentation_area	Area covered by segmentation
brightness_score	Brightness score
michelson_contrast_score	Michelson contrast score
bbox_x_min, bbox_y_min, bbox_x_max, bbox_y_max	Bounding box coordinates
class_name	Class name of the object
Model Generated	
Attribute	**Description**
average_depth	Average depth of the object
top_95_depth	Depth of the top 95% portion of the object
bottom_5_depth	Depth of the bottom 5% portion of the object

Table 4.3: **Overview of metadata sources used for enriching instance segmentation datasets.** Metadata was created from existing instance segmentation annotations and specialized models. The extraction is performed automatically without human input with heuristics and models.

Automatic Task Processing

To create a concrete list of vision-language tasks for each image we employed a systematic process. We began by prioritizing images in the datasets that featured a higher number of classes and objects to maximize task diversity and complexity. In the implementation, we sampled equally based on their instance segmentation count. Under regular object distributions and selection criteria, all images with a higher number of objects are selected. Next, specific criteria for each task were evaluated to ensure appropriate task generation for each image. For instance, in tasks requiring the comparison of two objects, it was essential that both objects were present in the image and belonged to the relevant classes. Furthermore, we established minimum thresholds for various measures, such as requiring a substantial depth difference between objects, to ensure the correct answers for the task could be reliably determined. We determined these thresholds with a human reader study (n=3). If future depth models generate more accurate absolute depth predictions, the threshold can be lowered accordingly. Overall, our objective was to generate as many of the 22 different tasks as possible for each image. No LLMs or VLMs were used for task generation, as these methods are prone to injecting hallucinations [199, 201]. We prioritized quality and reliability instead. Furthermore, our task processing pipeline is deterministic.

Human Ambiguity Baseline

To rate the difficulty and ambiguity for each of the 32,771 tasks, we further acquired annotations from six human raters per image. To use our monetary resources efficiently, we implemented

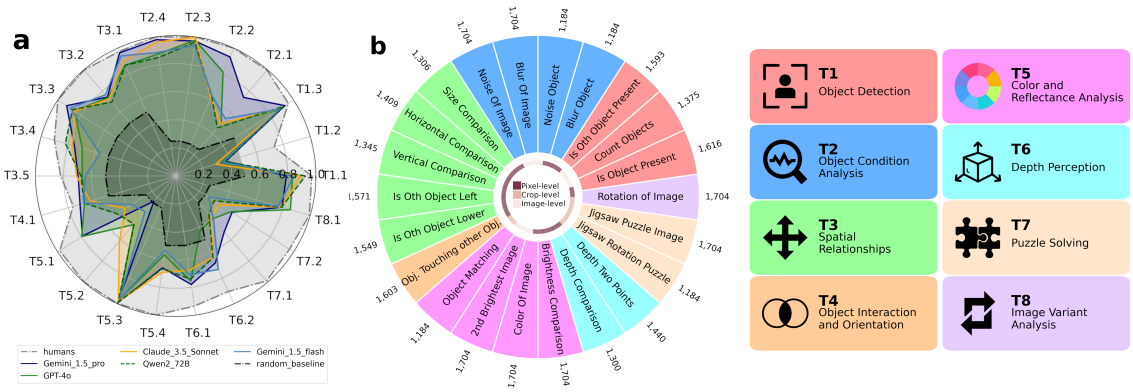


Figure 4.17: **DomainBench yields a diverse set of perception tasks on the image and object level.** (a) The spider diagram illustrates high Accuracy variability across tasks for the VLMs. We present the results of all the best ranked models while a comprehensive performance summary for all 22 tested models can be found in Appendix F.1. (b) Based on a single image with instance segmentations, DomainBench enables the generation of 22 tasks from eight different vision-language categories, ranging from pixel-level to image-level perception.

early stopping if four raters reached agreement on a task. Across the seven generated datasets, early stopping occurred in between 0.8% and 2.7% of the cases. Overall, this resulted in 142,333 human reference annotations. An overview of the resulting datasets is provided in Table 4.4 and exemplary images for all generated datasets are included in Appendix C.

Benchmarking Workflow

To ensure fair and consistent evaluation of all selected VLMs, we developed a standardized benchmarking workflow applied uniformly across all models. Future model We assessed them in a zero-shot setting without any additional fine-tuning or domain-specific training to probe their generalist abilities. We strictly followed the configurations and setups recommended by each model's authors, using the exact settings provided in their official repositories (e.g., on Hugging Face) to ensure that each model was evaluated under conditions intended by its creators. Each model was provided with a carefully crafted text prompt alongside the corresponding image. To eliminate potential ambiguities in the questions, we conducted iterative testing of these prompts among four human evaluators in our department. Through four rounds of refinement, we adjusted the prompts until all four human evaluators consistently agreed on their interpretation. We specifically did not tune our prompts with declarative frameworks, such as DSPy, to prevent advantages of single models in our current benchmarking effort. Optimizing the prompts for several included models would disadvantage both existing models not included in our benchmarking and future model releases. Furthermore, we evaluated the sensitivity of the VLMs to variations in image markers, as many questions involved marked objects. Altering the box colors used to highlight objects, from green and red to other colors, resulted in slight performance fluctuations in both directions across different VLMs. To maintain consistency, we used the commonly recognized colors red and green, assigning them to objects at random.

Domain	Icon	#Images	#Objects	#Tasks	#Human Baseline Annotations
Wildlife	🦁	268	853	5,528	24,024
Persons	🧍	250	7,812	6,122	26,548
Vehicles	🚗	235	2,199	5,219	22,976
Animals	🐕	273	1,162	5,724	24,907
Kitchen	🍳	272	2,143	5,332	23,793
Food	🥦	236	5,673	5,249	23,221
Kitti	🐱	170	1,458	3,997	17,477
Total		**1,704**	**21,300**	**32,771**	**142,333**

Table 4.4: **Dataset statistics across different domains.** The table presents the total number of images, objects, tasks, and human annotations across all domains.

Model	Overall	wildlife 🦁	animals 🐕	kitti 🐱	person 🧍	vehicles 🚗	food 🥦	kitchen 🍳
Humans	94.9 (±0.6)	95.7 (±0.6)	95.6 (±0.6)	95.2 (±0.7)	96.6 (±0.5)	93.2 (±0.7)	94.2 (±0.7)	93.5 (±0.7)
Gemini Pro	74.8 (±1.2)	77.6 (±1.2)	78.0 (±1.1)	79.5 (±1.3)	74.1 (±1.2)	74.9 (±1.3)	71.2 (±1.3)	68.2 (±1.3)
Claude_3.5_Sonnet	71.0 (±1.3)	76.9 (±1.2)	76.6 (±1.2)	74.0 (±1.5)	68.0 (±1.2)	70.7 (±1.3)	66.0 (±1.3)	64.9 (±1.3)
GPT-4o	70.8 (±1.3)	73.5 (±1.3)	73.0 (±1.2)	76.7 (±1.4)	70.1 (±1.2)	70.8 (±1.3)	66.8 (±1.3)	64.3 (±1.4)
Qwen2_72B	70.3 (±1.3)	72.7 (±1.3)	76.5 (±1.2)	75.5 (±1.4)	63.9 (±1.3)	70.3 (±1.3)	67.3 (±1.3)	65.6 (±1.3)
Llama_3.2_90B	67.5 (±1.3)	73.5 (±1.3)	72.1 (±1.2)	70.8 (±1.5)	65.0 (±1.3)	66.2 (±1.4)	63.3 (±1.4)	61.7 (±1.4)
Gemini Flash	67.5 (±1.3)	74.5 (±1.2)	71.7 (±1.3)	71.1 (±1.5)	62.8 (±1.3)	67.9 (±1.4)	62.3 (±1.4)	62.2 (±1.4)

Table 4.5: **The rankings of models differ strongly across the tested domains.** Model Accuracies across different generated datasets (±95% confidence interval). The 'Overall' column represents the mean accuracy across all datasets. 1st place (Gold) 2nd place (Silver) 3rd place (Bronze) 4th place 5th place 6th place. Only the top six models are shown. The 'Overall' column represents the mean accuracy across all datasets. Due to space constraints, results for additional models are provided in Table A.4. Note that Accuracy does not account for shared images between questions; this issue is addressed in Figure 4.19.

Metrics and Rankings

Choosing an adequate strategy for performance assessment is far from trivial and a research topic of its own [118, 161]. In this work, we were specifically interested in relative performance differences rather than in the specific ability of VLMs to serve a specific task. A major drawback of accuracy is that it treats each question independently and does not account for shared structure, overlap, or dependencies across questions (here the shared base image for generating n questions). To obtain aggregated performance values across images, we define the **Accuracy%(t)** metric with a threshold $t \in [0, 1]$. For each image i in a dataset D, let Q_i denote the set of questions associated with that image. Let $C_{i,q,m} \in \{0, 1\}$ indicate whether model m correctly answered question q for image i (1 for correct, 0 otherwise). The model m is considered to meet the threshold t on image i if the fraction of questions q in Q_i answered correctly by the model is at least t. Formally, we define:

$$\text{Accuracy\%}_m(t) = \frac{1}{|D|} \sum_{i \in D} I\Big(\big(\tfrac{1}{|Q_i|} \sum_{q \in Q_i} C_{i,q,m} \big) \geq t \Big) \times 100 \tag{4.5}$$

Here, $I(\cdot)$ is an indicator function defined as:

$$I(x \geq t) \; = \; \begin{cases} 1, & \text{if } x \geq t, \\ 0, & \text{otherwise.} \end{cases} \tag{4.6}$$

Explanation:

- $\sum_{q \in Q_i} C_{i,q,m}$: Total number of correctly answered questions for image i.
- $\frac{1}{|Q_i|} \sum_{q \in Q_i} C_{i,q,m}$: Fraction of questions answered correctly for image i.
- $t \in [0, 1]$: Desired minimum accuracy level assessed for each Q_i.

4.3.3 Experiments

In the following, we describe the experimental setup for quantifying the performance and usability of our proposed framework DomainBench, the generated datasets and our large-scale study on SOTA VLMs. VLM benchmarking results can vary substantially with various factors, such as the images used, the domain, and the applied prompts. This often renders comparison of results across papers infeasible. To address this bottleneck, we fully homogenized our benchmarking framework using the proposed framework.

VLM Tasks Selection

We evaluated the selected models on a comprehensive set of 22 diverse tasks derived from our task augmentation framework DomainBench (overview in Figure 4.17, full list in Appendix B.2 and examples per dataset in Appendix C). Each task was associated with specific evaluation criteria and standardized prompts, as outlined in the Methods section. For instance, when dealing with multiple-choice questions or tasks involving object selection, we established clear guidelines on how options were presented and how objects were chosen within images. For the object selection we first generate all possible combinations within the defined thresholds and then select the final combination at random. This attention to detail ensured that the evaluation was both rigorous and reproducible. Further details can be found in the code and users can adjust all thresholds according to their domain requirements.

Model Selection

We selected 22 frontier and open VLMs of various sizes and from various providers and sources, as illustrated in Table 4.6. The oldest model was released in January 2024, while the most recent one included was released at the end of September 2024. This includes closed source models from the leading labs OpenAI, Anthropic, Google Deepmind and a selection of open source models with varying sizes. Since the closed source models provide no information about the model size, we can not include this factor in our analysis.

Access	Size	Name	Version	Organization	Release Date
Closed	-	GPT-4o	gpt-4o-2024-08-06	OpenAI	2024-08-08
Closed	-	GPT-4o-mini	gpt-4o-mini-2024-07-18	OpenAI	2024-07-18
Closed	-	Gemini 1.5 Pro	gemini-1.5-pro-001	Google	2024-05-24
Closed	-	Gemini 1.5 Flash	gemini-1.5-flash-001	Google	2024-05-24
Closed	-	Claude 3.5 Sonnet	claude-3-5-sonnet-20240620	Anthropic	2024-06-20
Open	1B	InternVL2-1B	InternVL2-1B	OpenGVLab	2024-07-04
Open	8B	InternVL2-8B	InternVL2-8B	OpenGVLab	2024-07-04
Open	40B	InternVL2-40B	InternVL2-40B	OpenGVLab	2024-07-04
Open	7B	Qwen2 7B	Qwen2-VL-7B-Instruct	Alibaba	2024-08-30
Open	72B	Qwen2 72B	Qwen2-VL-72B-Instruct	Alibaba	2024-08-30
Open	7B	LLaVA-NeXT 7B	llava-v1.6-mistral-7b-hf	U. of Wisconsin-Madison	2024-01-30
Open	34B	LLaVA-NeXt 34B	lava-v1.6-34b-hf	U. of Wisconsin-Madison	2024-01-30
Open	7B	Chameleon 7B	chameleon-7b	Meta	2024-05-16
Open	4.2B	Phi-3 Vision	Phi-3-vision-128k-instruct	Microsoft	2024-04-23
Open	4.2B	Phi-3.5 Vision	Phi-3.5-vision-instruct	Microsoft	2024-08-20
Open	770M	Florence-2	Florence-2-large-ft	Microsoft	2024-06-15
Open	3B	PaliGemma 3B 224x224	paligemma-3b-mix-224	Google	2024-05-14
Open	3B	PaliGemma 3B 448x448	paligemma-3b-mix-448	Google	2024-05-14
Open	12B	Pixtral	Pixtral-12B-2409	Mistral	2024-09-17
Open	11B	Llama 3.2 11B	llama-3-2-11b-vision-instruct	Meta	2024-09-25
Open	90B	Llama 3.2 90B	llama-3-2-90b-vision-instruct	Meta	2024-09-25
Open	7B	Molmo 7B	Molmo-7B-D	Allen Institute for AI	2024-09-24

Table 4.6: **Our 22 evaluated SOTA vision-language models span both open and closed-source architectures.** The benchmark includes 22 VLMs with precisely specified versions, representing current capabilities across both proprietary and publicly available models. This includes closed source models from the leading labs OpenAI, Anthropic, Google Deepmind and a selection of open source models with varying sizes.

Seven New Datasets from Diverse Domains

We applied our proposed framework DomainBench to images from seven different domains. Overall, the input images and instance segmentations for our framework were extracted from KITTI [57], COCO [109], and COCONut [40]. In summary, we added ~300,000 metadata annotations to a total of 1,704 images across seven domains, ranging from kitchen environments to animals. This includes 12 annotations per object (e.g., relative size, touching segmentation masks, or relative depth). The complete list is provided in Table 4.3.

The metadata were then used to define a set of 22 different VLM tasks (see Figure 4.17), including six tasks concerning the entire image, ten related to individual objects, and six focused on object pairs. For further detail we refer the reader to the Methods section.

4.3.4 Results

The primary purpose of our experiments was to showcase the benefit of our task augmentation approach (Section 4.3.4). To assess the value of each task for VLM benchmarking, we related it to average model performance, resources needed to create the task, and corresponding human

ID	Task Name	Task Description	Answer Type
T1.1	Is Object Present	Determines whether a specified object is present in the image.	Binary
T1.2	Count Objects	Determines the number of objects in the image	Count
T1.3	Is Oth Object Present	Determines whether or not there is more than one object in the image	Binary
T2.1	Blur Object	Determines whether an object is blurred	Quiz (A/B/C/D)
T2.2	Noise Object	Determines whether an object contains noise	Quiz (A/B/C/D)
T2.3	Blur Of Image	Determines which image variant is least blurred	Quiz (A/B/C/D)
T2.4	Noise Of Image	Determines which image variant is not corrupted	Quiz (A/B/C/D)
T3.1	Size Comparison	Determines which of two objects is larger	Color
T3.2	Horizontal Comparison	Determines which object is further to the left of the image	Color
T3.3	Vertical Comparison	Determines which object is further to the bottom of the image	Color
T3.4	Is Oth Object Left	Determines whether there is another image further to the left of an object	Binary
T3.5	Is Oth Object Lower	Determines whether there is another image further to the bottom of an object	Binary
T4.1	Is Object Touching other Object	Determines if two objects are touching each other	Binary
T5.1	Color Object Matching	Determines which of four tiles show the correct color for the given image	Quiz (A/B/C/D)
T5.2	2nd Brightest Image	Determines which of the images is the 2nd brightest image	Quiz (A/B/C/D)
T5.3	Color Of Image	Determines which image variant is not corrupted	Quiz (A/B/C/D)
T5.4	Brightness Comparison of Two Points	Determines which of two points is brighter	Binary
T6.1	Depth Comparison	Determines which of two objects is closer to the camera	Color
T6.2	Depth Two Points Image	Determines which point is closer	Binary
T7.1	Jigsaw rotation Puzzle	Determines which of four rotated tiles fits best into a cut out area of the image	Quiz (A/B/C/D)
T7.2	Jigsaw Puzzle Image	Determines which of four tiles fits best into a cut out area of the image	Quiz (A/B/C/D)
T8.1	Rotation Of Image	Determines which image variant is not rotated	Quiz (A/B/C/D)

Table 4.7: **Overview of VLM Benchmark Tasks generated with the framework.** We provide a small task description and answer type for each generated task. Examples across datasets are displayed in subsection C.1.

ambiguity (Section 4.3.4). Finally, we leveraged our concept and data to explore the capabilities of the most recent open source and closed source VLMs (Section 4.3.4.).

Dataset Publications Are Becoming the Norm

CVPR 2024		Manually verified papers	Agreement LLMs and Human verification
Total number of papers	2,708	–	–
With new or modified dataset:	397	40 (10%)	1
Without new or modified dataset:	2,311	50 (2%)	1

Table 4.8: **A notable portion of CVPR 2024 papers contribute new or modified datasets, highlighting a rising trend in dataset-focused research**. CVPR 2024 paper analysis summary.

Numerous datasets and benchmarks are continually being released in the general computer vision field. According to our analyses, ∼400 out of the 2,700 CVPR 2024 publications propose a new or modified dataset as detailed in Table 4.8. These benchmarks cover a wide range of domains, from autonomous driving to wildlife monitoring, underscoring the need for *domain-specific* benchmarks. A large fraction of these datasets contains a high percentage of annotation errors, as displayed in Figure 4.18.

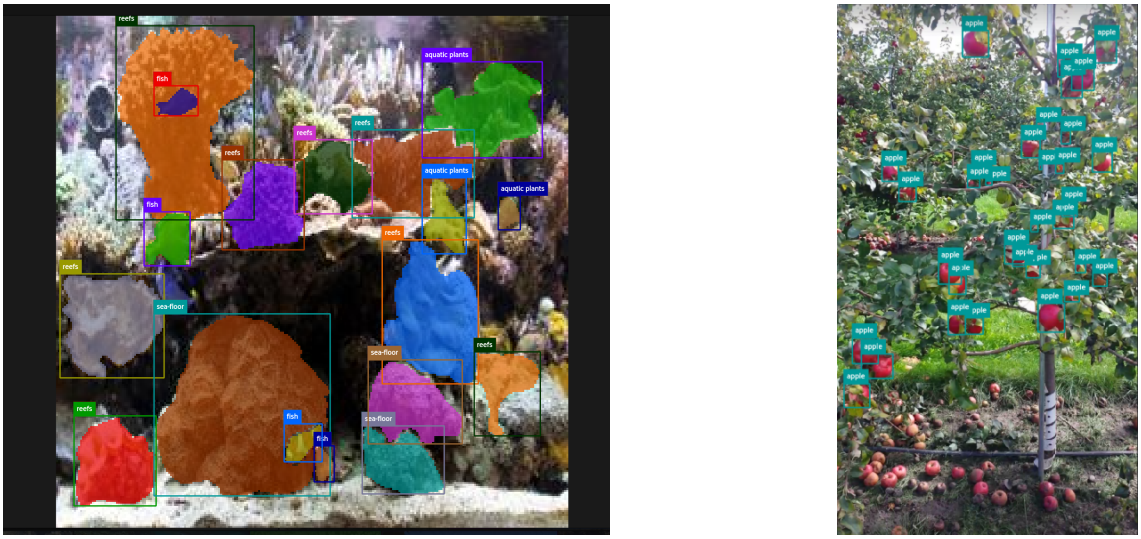


Figure 4.18: **A large fraction of newly released datasets at major computer vision conferences contains a high percentage of annotation errors.** The errors were identified within 100 seconds of loading the dataset. (left) Several fish, reefs and water plants are not annotated. Image from [106]. (right) Several apples on the tree, the neighboring tree and the floor are not annotated. Image from [63].

Benefit of The Proposed Framework

Table 4.5 shows aggregated performance values for the top six models, separated by imaging domain. As the tasks and prompts were homogenized, the results clearly indicate that performance

on the same question tasks varies substantially across domains, supporting the hypothesis that in-domain validation is crucial for real-world translation. Note that this holds true despite the fact that we purposely chose domains that are relatively common (presumably captured in the model training) and closely related to one another. While Gemini Pro achieves the strongest performance across all seven datasets, the gap to human performance remains between 18 to 25 absolute points. We refer the reader to Section 4.4 for results on surgical imaging data.

Furthermore, as shown in Figure 4.17a, the performance of models varies substantially across VLM tasks, suggesting that the tasks generated by our framework are diverse. The hardest tasks on average across domains are (1) T7.2 “Jigsaw Puzzle Completion”, (2), T1.2 “Object Counting”, (3), T7.1 “Rotated Jigsaw Puzzle Completion”, and (4) T5.2 “Second Brightest Image Selection”. The easiest task on average was T1.3 “Additional Object Presence Detection” (see Figure 4.20).

Human Ambiguity

As demonstrated in Figure 4.21 there is a high discrepancy in task rankings between humans and models. While the "Jigsaw Puzzle Completion” tasks ranked amongst the most challenging for the models, humans found "Object Counting" and “Object Touching Detection” to be the most difficult.

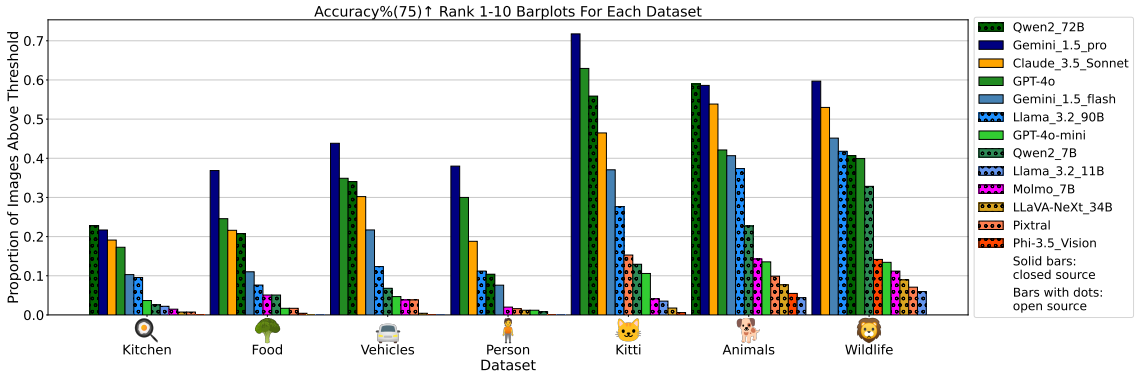
From a resource perspective, tasks should be (1) hard to solve for models and (2) require as little human annotation as possible. This potential trade-off is captured in Figure 4.20. It can be seen that many hard tasks generated by DomainBench, including the top four, can already be extracted from instance segmentations alone. This pattern can be observed across tasks that rely on instance segmentation, whole-image input, or external depth models. The lowest human agreement occurs on the object counting task, likely because several images in the dataset contain a high number of objects. In contrast, several tasks yield near-perfect human agreement, while ã third of the models fails. For a full overview of task rankings we refer the reader to Appendix E.2.

Insights on Current Vision-Language Models

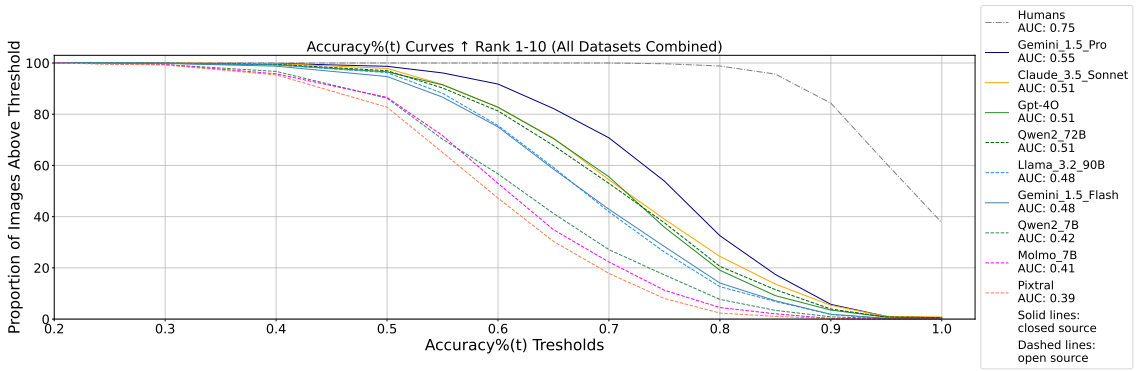
Figure 4.19 summarizes the performance of a model selection and reference baselines. Further detailed analysis, including all tested models, examples, and errors for each generated dataset are provided in the Appendix. The following insights can be extracted:

Confirming common findings from the community:

Our analysis reinforces several established patterns in the VLM field. Closed models continue to demonstrate superior performance across tasks and domains, although open models have significantly narrowed this performance gap. In particular, Qwen2 72B [206] stands out as the strongest performer among open models across our experiments. The superiority of human evaluation remains evident, with human raters achieving near-perfect performance on most tasks, though they notably struggle with specific challenges such as counting, and object touching-related tasks, counting being particularly problematic. Regarding model scaling, larger variants typically show better performance across all tasks. Notable exceptions include Molmo 7B [37] outperforming Pixtral 12B with 40% less parameters.



(a) **The need for specific in-domain evaluation is demonstrated by the high performance variability across imaging domains.** The performance of the overall best model Gemini Pro varies between domains from 22% (Kitchen dataset) to 72% (Kitti dataset). For the displayed Accuracy%(75), humans achieve an almost perfect score of 1 for all datasets (see Appendix). The top 10 models per dataset shown. We display the full plots for all thresholds and models in Appendix D.1.



(b) The Area under the Accuracy%(t) Curve serves as a **metric for comprehensive image understanding**. With a maximum possible value of 1, higher values indicate better performance. Notably, the current SOTA model, Gemini Pro, achieves only 0.55, highlighting significant room for improvement. Only the top 10 models are shown. Separate curves for all 22 tested models and datasets are displayed in Appendix D.3.

Figure 4.19: **Performance varies across domains, highlighting the need for specialized in-domain evaluation; even the best models still lag behind human performance**. The Accuracy%(t) metric represents the percentage of images for which at least a specified proportion of questions are correctly answered. It can (a) be computed for specific thresholds or (b) be aggregated over multiple thresholds to remove dependence on a specific t. The Area under the Accuracy%(t) Curve captures model performance in a single value, ranging from 0.39 to 0.55 for the top 10 models tested.

Interesting new findings:

The need for specific in-domain evaluation is highlighted by the high performance variability across imaging domains for the same perception tasks, see Table 4.5 and Figure 4.16. The overall best model, Gemini Pro, varies between domains from 22% (Kitchen dataset) to 72% (Kitti dataset). Qwen2 72B [206] slightly surpasses Gemini Pro on the kitchen and animals datasets but ranks only fifth on the person dataset.

Additional insights emerge from pairwise model comparisons, with Qwen2 7B consistently outperforming Molmo 7B [37] across most datasets, indicating that the Molmo training strategy may have removed certain capabilities. Molmo 7B is built upon Qwen2 7B with the PixMo dataset [37]. The smaller Gemini variant, Gemini Flash, exhibits superior Point Depth Comparison capabilities compared to Gemini Pro. This suggests that the two models were trained with different strategies, since a simple distillation process would typically be expected to reduce performance rather than enhance it. The overall results indicate that our newly introduced metric, Accuracy%(t), can effectively capture model performance across tasks for the same image in a single value. Further analysis reveals specific weaknesses even in top-performing models (visualizations available in Appendix F.1). For instance, Gemini Pro struggled notably with the 'Second Brightest Image' task (T5.2, Acc: 23.6%, ranking 19th compared to Claude 3.5 Sonnet's 61.1% at rank 1st). GPT-4o showed relative weakness on 'Point Depth Comparison' (T6.2, Acc: 43.3%, rank 15th) and 'Additional Object Presence Detection' (T1.3, rank 11th). Even the 5th-ranked Llama 3.2 90B performed poorly on 'Object Blur Detection' (T2.3, Acc: 30.2%) and 'Object Noise Detection' (T2.4, Acc: 24.8%), despite ranking first on horizontal and vertical spatial reasoning tasks (T3.2, T3.3). Llama 3.2 11B struggled with 'Color Object Matching' with only 7.7%, where even the small InternVL2 1B obtained 23.4%. In contrast to these specific capability gaps, human annotator errors primarily occurred on inherently ambiguous or visually challenging samples, such as objects partially hidden in shadows, entering the frame abruptly, or positioned at ambiguous mid-range depths."

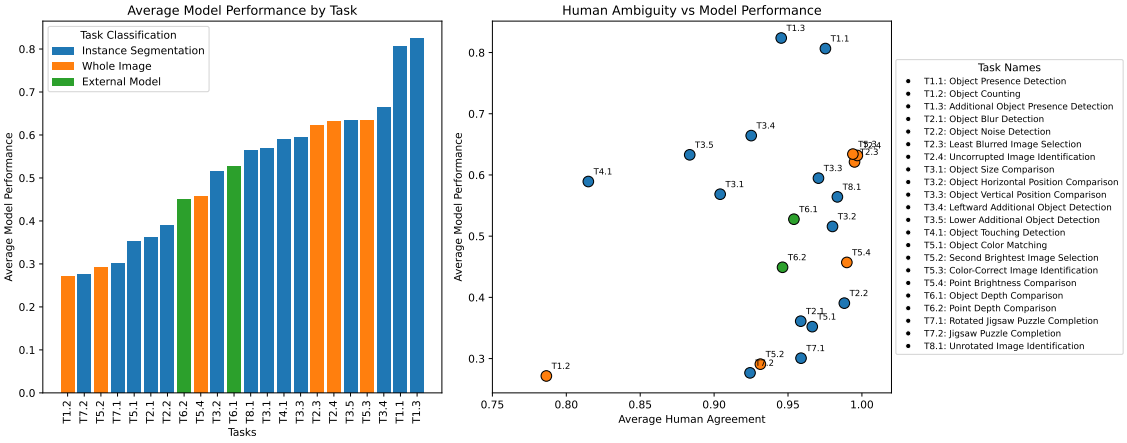


Figure 4.20: **Instance segmentations alone allow for the extraction of hard tasks.** (a) Tasks were classified in those extractable directly from instance segmentations (blue), requiring external models (green) and requiring the whole image (orange). DomainBench is able to automatically extract a broad range tasks that challenge SOTA VLMs. (b) Human ambiguity plotted against model performance. Humans are able to generate near perfect agreement for a large part of the tasks. Only two tasks obtain an agreement below 0.85.

4.3.5 Discussion

This paper contributes to the advancement of VLM benchmarking in three ways:

1) DomainBench, a framework for resource-efficient and domain-specific benchmarking: We showed that task augmentation, using instance segmentation as the root task, enables the generation of a diverse set of VLM tasks and could thus evolve as a core method for resource-efficient domain-specific VLM benchmarking. The insights gained on the varying difficulty of presented VLM tasks will further guide the design of future benchmarks. The framework can be easily applied to other domains, even with a small number of images. The computational and monetary costs for each generated dataset are minimal and displayed in Appendix B.3. This makes it an excellent choice for domain researchers to probe current VLMs capabilities on their data.

2) Seven new openly available datasets: Our seven new datasets will help assess generalist capabilities of future VLMs. They cover a diverse range from kitchen scenes to animals. Furthermore, we release the six human annotations per task (totaling 142,333 annotations) to assist researchers working on human annotations.

3) New insights: The insights on current capabilities of closed and open VLMs highlight the narrowing gap between closed and open models. Most importantly, we showcased the need for domain-specific validation. Furthermore we display family specific weaknesses across the 22 tested models. This indicates that the training strategies or training data differs between the major labs and users should select their VLM of choice with care.

Core strengths of our contribution include the broad applicability of our concept, the open dataset and benchmark contribution, and the wide range of SOTA closed and open models investigated here.

As an implicit contribution, we introduced the new metric Accuracy%(t), which offers several key strengths. First, it captures model performance in a single very intuitive value. The metric is extensible with additional tasks, allowing for gradually increasing difficulty, and can be adapted to evaluate domain-specific tasks effectively. It is worth mentioning, however, that the specific properties of the metric require further analyses [161]. For example, some questions require specific image conditions, such as the presence of multiple objects for comparison. This can result in a varying number of questions per image, which, in turn, has an influence on the metric. Furthermore, tasks are treated equally without any weighting, which may overlook differences in task difficulty or importance. Users can, however, easily modify the weighting scheme to better reflect their specific evaluation priorities. Similarly, users can exclude irrelevant questions from the benchmark to safe inference costs and tailor the benchmark to their own requirements.

To understand relationships between the generated tasks, we computed Pearson correlations between task accuracies across the top five models for each of the seven datasets, as displayed in Appendix F.2. Generally, correlations were weak across the different domains, with one consistent exception: tasks involving object modifications (e.g., jigsaw puzzles, recoloring, noise) formed a coherent cluster, suggesting they assess similar underlying model capabilities. Other potential task relationships proved highly domain-dependent. For example, vertical and horizontal comparison tasks correlated strongly within the Animals dataset (r=0.973) but showed much lower average correlation elsewhere (avg r=0.336). This finding further underscores the necessity for domain-specific benchmarking, as task relationships and underlying skill requirements can vary significantly with the visual context.

We also investigated the impact of image resolution, a known factor influencing VLM performance. Downscaling a 25% subset of the Animal dataset images (original width: 320-1032 px, height: 194-660 px) and re-evaluating with the top model (Gemini Pro) showed an accuracy reduction of 4.4% when scaling to 224px and a more significant drop of 17.2% when scaling to 112px. This suggests, as expected, that utilizing higher native image resolutions likely yields better performance on these perception tasks.

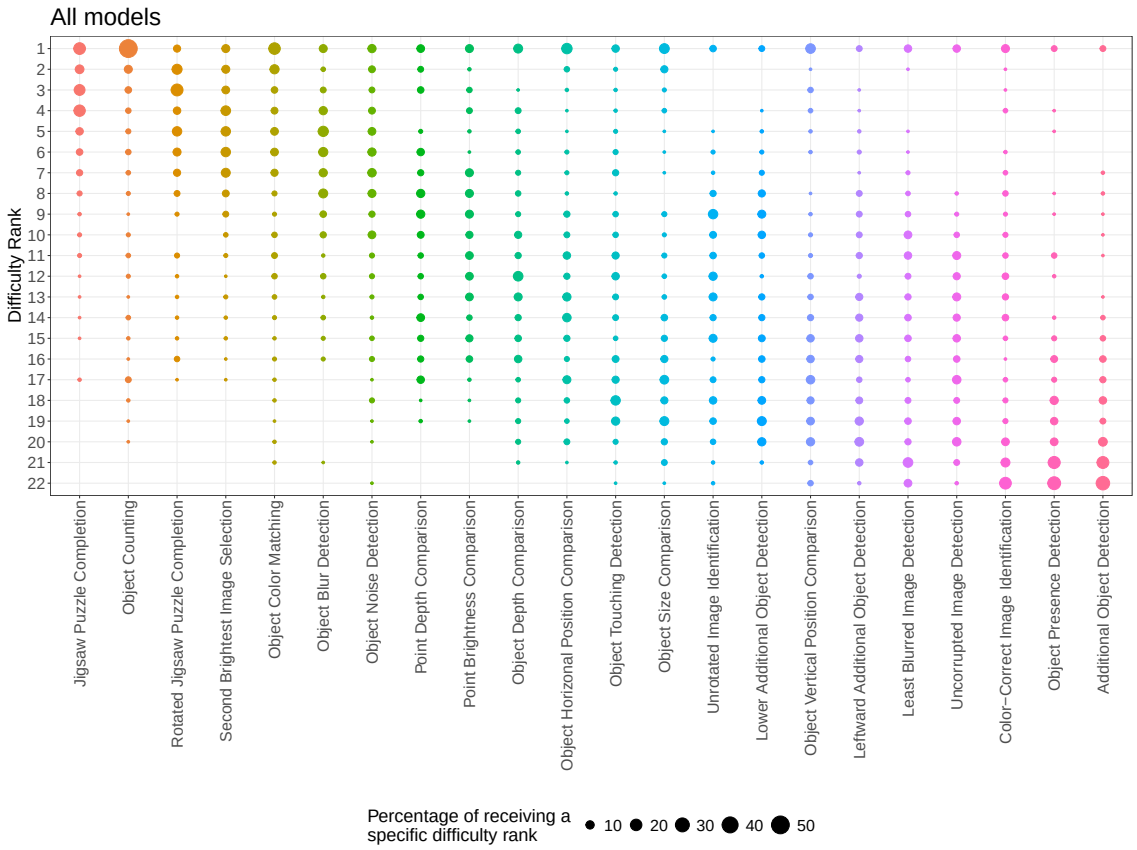


Figure 4.21: **All tested VLMs, both open and closed-source, exhibit consistent patterns in task difficulty.** The plots show the difficulty of tasks based on aggregated model scores/human scores (1 = hardest task, 22 = easiest task). The radius of the blob indicates how often a task was assigned a difficulty rank when considering all seven domains and all models (n = 5 for closed models; n = 16 for open models; n = 21 for all models; n = 1 for humans as majority vote over several raters). The larger the plot, the higher the percentage it achieved a specific rank. The hardest tasks on average across domains are (1) T7.2 “Jigsaw Puzzle Completion”, (2), T1.2 “Object Counting”, (3), T7.1 “Rotated Jigsaw Puzzle Completion”, and (4) T5.2 “Second Brightest Image Selection”. The easiest task on average was T1.3 “Additional Object Presence Detection”.

A limitation of our work is model family dependence, as many models come from closely related families, which may hinder statistical analysis. For closed-source models, specific information about training and data is often unavailable, creating transparency issues. This is further substan-

tiated by our results with the Gemini family that indicate the current Gemini Flash is not just a distilled version of Gemini Pro. As an ablation experiment addressing output variability, we re-ran 71 randomly selected questions across 8 models 10 times each; all answers remained identical, confirming minimal output variability under the experimental setting (temperature zero and fixed seeds). We provide further statistical analysis, such as ranking variability in Appendix E. Model performance showed small variations with prompt phrasing, which we mitigated through iterative testing for consistency. However we do note that prompt phrasing can have a substantial effect on model performance. Additionally, our human annotations were performed by professional annotators, which may introduce ambiguity since annotators aim to complete tasks quickly to increase their hourly pay.

Future work should focus on expanding the number of tasks generated, further enhancing the diversity and comprehensiveness of VLM benchmarks. Additionally, our method can be adapted to different domains with domain-specific questions or scaled up to support continuous extension, providing a versatile approach for evaluating models across diverse applications. Furthermore it should be applied to new models to monitor their rising capabilities until the benchmark can be considered solved.

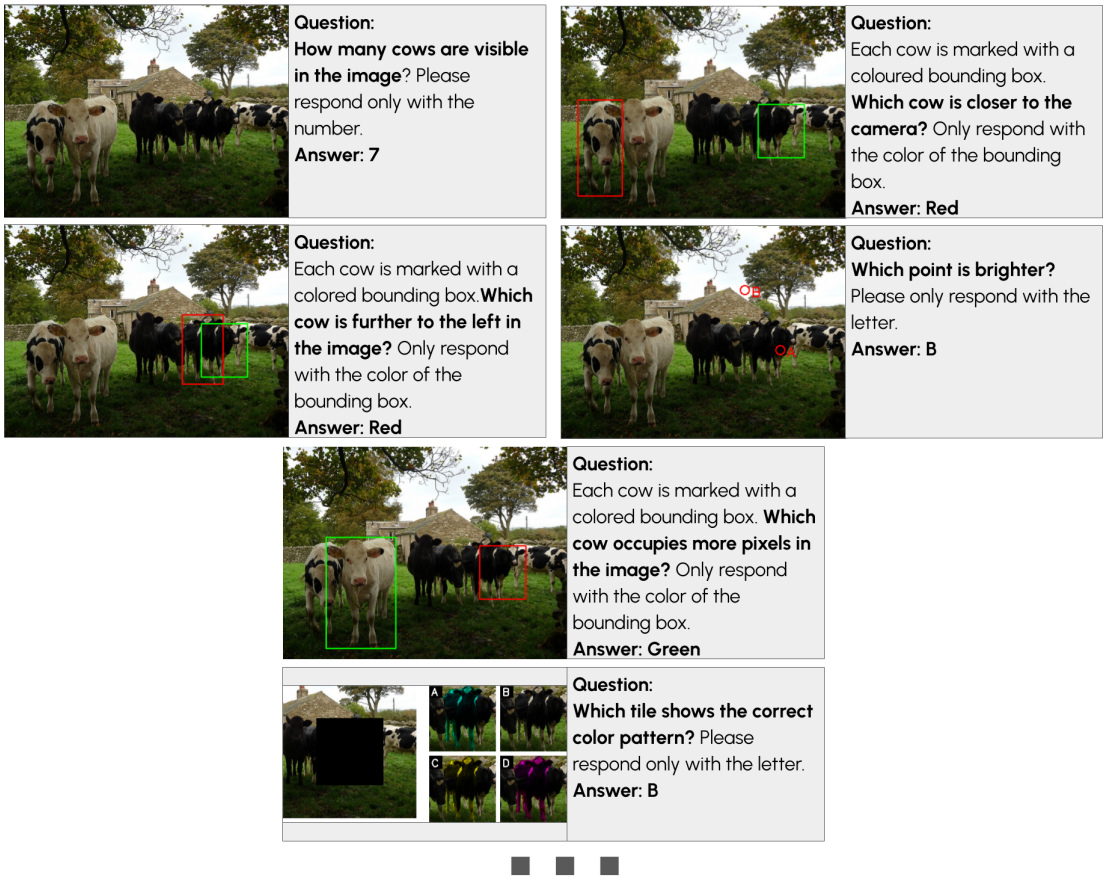


Figure 4.22: **DomainBench yields a diverse set of 22 tasks.** Exemplary tasks that were generated with the framework for a given image with the instance segmentation, the whole image and information provided by external models. A broad range of examples and errors for each generated dataset is provided in Appendix C. The tasks cover the capabilities on then object and the image level.

4.4 Applying DomainBench to Surgical Data

This section is based in large parts on the submitted manuscript: Leon Mayer, Tim Rädsch et al., "Challenging Vision-Language Models with Surgical Data: A New Dataset and Broad Benchmarking Study", submitted to Medical Image Analysis in 2025 (preprint: [122]). As the core contributor alongside Leon Mayer, I played a central role in designing the study, preparing and conducting the annotation experiments, generating the benchmark datasets and implementing the associated code. I initiated and obtained the funding for the collaboration with the involved QA company and annotators. This includes the contractual agreements, general setup and code for interacting with the platform. I led the creation of the data enrichment pipeline and the benchmark pipeline. Moreover, I was the principal creator of seven out of the eight datasets utilized in the study. This included not only the construction and curation of these datasets but also the preprocessing and data format of all eight datasets to ensure they met the standards required for the benchmarking pipeline. I generated the main figures of the work. More specifically, I was responsible for generating Figure 1 and Figure 2, which are central to the presentation of our results. I contributed substantially to the writing and discussion of the paper. Significant portions of the text, figures, and results have been adapted and integrated into this chapter. Text and figures are extracted from the preprint at the time of submission. The concept passage has been rewritten so that each experiment's rationale, finding, and implication is clearly explained, and the respective Discussion subsection enables readers to understand both the integrated contribution and its individual ramifications. This work was co-authored with Leon Mayer, Dominik Michael, Lucas Luttner, Amine Yamlahi, Evangelia Christodoulou, Patrick Godau, Marcel Knopp, Annika Reinke, Fiona Kolbinger, Lena Maier-Hein.

4.4.1 Concept Overview

Rigorous validation of VLMs in surgical imaging builds the foundation for both scientific progress and clinical translation. While traditional computer vision models have historically struggled to generalize to endoscopic domains [117], FMs have demonstrated improved cross-domain transfer. Yet, the capabilities of VLMs for endoscopic tasks remain largely under- or even unexplored.

We identify three critical gaps motivating our systematic study:
Heterogeneous datasets hinder comparison:
Existing endoscopic VQA datasets vary widely in question style and task scope, preventing alignment with findings observed on natural images.
Lack of large-scale assessments on endoscopic data:
Prior endoscopic VQA studies rely on small, domain-specific models, with no comprehensive evaluation of SOTA VLMs.
Unclear role of medical versus generalist models on endoscopic data:
No direct comparisons exist between size-matched generalist and medical VLMs, despite indications that generalist models may outperform their in domain variants.

To address these gaps, this work introduces three main contributions: (1) the first large-scale analysis of VLMs in laparoscopic surgery across basic perception and advanced frame-based tasks; (2) the release of HeiCo-VQA, a new dataset of 24,252 images and 167,384 questions with human baselines for 10% of samples, and (3) a comparative benchmark of several SOTA VLMs, both generalist and medical, on surgical data.

Our results reveal that (i) VLMs achieve strong performance on basic perception tasks such as object counting and localization, comparable to general domain tasks (ii) performance deteriorates substantially for tasks requiring in-depth medical knowledge, and (iii) specialized medical VLMs currently underperform their generalist counterparts across both task types. These findings underscore the need for improved modeling strategies fine-tuned to the unique challenges of surgical environments and provide a roadmap for developing next-generation endoscopic AI systems.

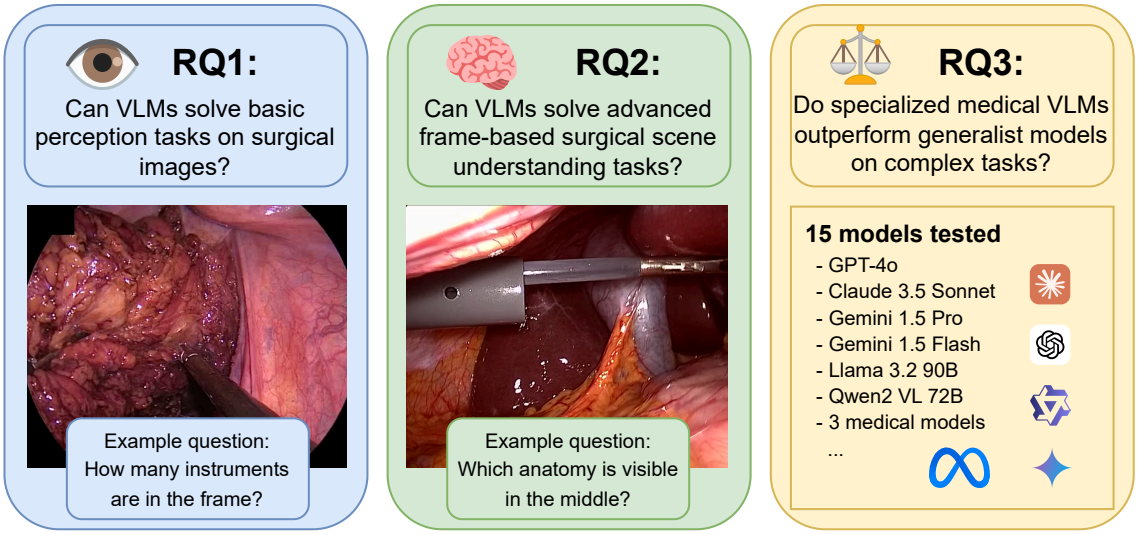


Figure 4.23: **Research questions addressed in this study** and sample images illustrating the nature and complexity of basic perception and advanced tasks.

4.4.2 Methods

This section introduces the framework for VQA benchmarking (Section 4.4.2), the new surgical VQA benchmark we release as part of this work (Section 4.4.2) as well as the specific experiments performed to address our core RQs(Section 4.4.3).

Framework for VQA Benchmarking

To investigate the RQs depicted in Figure 4.23, we developed a flexible framework that enables integration of any SOTA VLM while assessing performance across varying levels of complexity.

Basic Perception Tasks on Surgical Images

These elementary tasks evaluate a model's fundamental visual understanding without requiring any medical knowledge, thereby enabling cross-domain comparability. We convert a dataset of instance segmentations into a VQA dataset using the following pipeline and framework adopted from [156]. First, each image and its objects are enriched with relative depth information via the Depth Anything V2 model [208]. Next, a set of rules transforms object attributes, including depth, size, color, and spatial relationships of surgical instruments, into a set of facts (e.g., "there

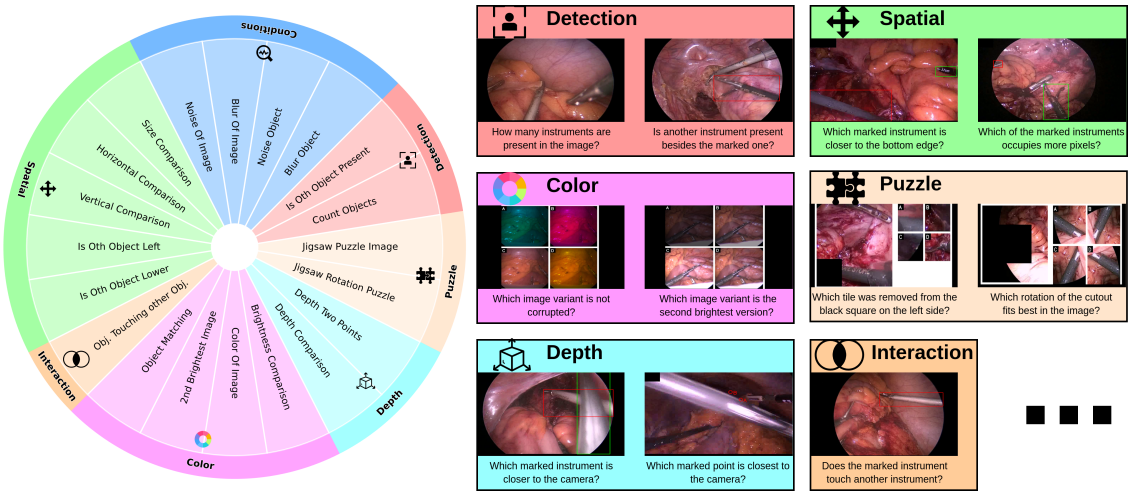


Figure 4.24: **The basic perception tasks cover a broad range of question categories.** We display sample questions for six categories, covering detection, spatial reasoning, color, puzzle, depth, and interaction questions. Example questions rephrased for brevity.

are N instruments in the image"). Following established VQA categorization [52], the resulting questions are subdivided into seven categories, depicted in Figure 4.24. All random baselines in our work are computed by selecting answers uniformly at random.

Advanced Frame-based Visual Scene Understanding

To estimate performance on more **complex tasks** requiring surgical knowledge and advanced reasoning, we convert existing endoscopic vision tasks into VQA formats. Classification tasks are straightforwardly rephrased into multiple-choice questions, enabling evaluation of nuanced surgical decision-making. Our framework seamlessly integrates any model that works with Huggingface Transformers [204] or is accessible via APIs, ensuring broad compatibility with SOTA VLMs. Performance is evaluated using two primary metrics. The **Accuracy(%)** metric [156] assesses model performance on image-question pairs by calculating the fraction of correctly answered questions per image; an image is deemed successful if this fraction meets or exceeds a specified threshold (e.g., 75%). Hence, a metric value of 0.5 at threshold 0.75 means that in 50% of images at least 75% of questions have been answered correctly. Additionally, **Matthew's Correlation Coefficient** provides a balanced measure of classification performance, a value of zero indicates performance equivalent to a random classifier.

Table 4.9: **Overview of the utilized endoscopic datasets.** These cover a broad range from general perception tasks to very hard endoscopic tasks.

Dataset	Specialty	Questions	Perception	Endoscopic Complexity			
				Simple	Medium	Hard	Very Hard
HeiCo-VQA	Rectal Surgery	167,384	✓	-	-	-	-
SSG-VQA	Cholecystectomy	883,254	-	✓	✓	✓	-
Kvasir-VQA	Endoscopy	58,798	-	✓	✓	✓	-
Endoscapes-CVS	Cholecystectomy	7,643	-	-	-	-	✓

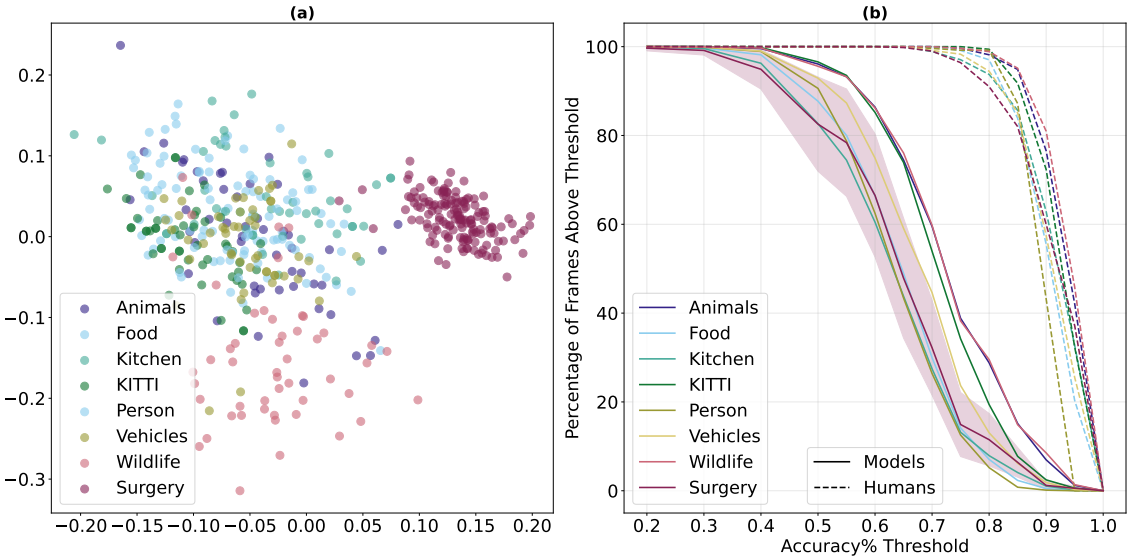


Figure 4.25: **Basic perception tasks are solved with comparable accuracy on surgical and natural images.** (a) PCA Visualization of the Qwen2 VL 7B vision encoder showing that surgical images (red) are encoded differently from natural ones. (b) Despite this, SOTA models perform similarly when solving basic perception tasks on the different domains. Each curve represents the average performance across models according to the Accuracy%(t) metric, which provides the percentage of images for which at least a specified proportion of questions (on the y-axis) are correctly answered. For the surgery data, bands represent the standard deviation (±) across the models. Performance of a random classifier is dataset-dependent, but generally close to zero.

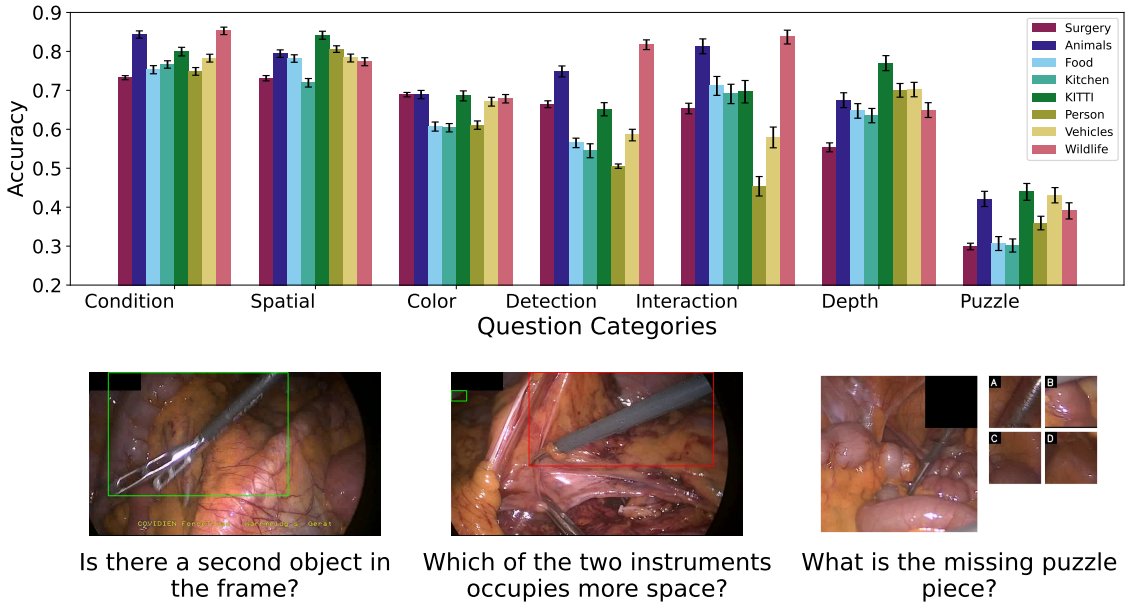


Figure 4.26: **Vision-Language models struggle on similar tasks in surgery as in other domains.** Performance on various task categories summarized in Figure 4.24 and representative sample questions. Models struggle especially with the puzzle question on the right. The error bars of each individual bar, correspond to the 95% confidence intervals that were calculated by performing bootstrapping for each question category using 10,000 resamplings.

New Dataset HeiCo-VQA

We created a new surgical VQA benchmarking dataset for basic visual perception tasks by applying our framework to the existing HeiCo dataset [116]. HeiCo comprises 30 laparoscopic surgical videos (10 each from proctocolectomy, rectal resection, and sigmoid resection procedures) with 10,040 frames containing instance segmentations. For each frame, we generated an average of 17 questions across all 7 categories depicted in Figure 4.24, resulting in 167,384 question-answer pairs. To establish a human baseline, we obtained reference answers (3 to 6 with early stopping based on human agreement) through annotations outsourced to QualityMatch GmbH, Heidelberg. In total, human reference annotations were obtained for 16,653 images.

4.4.3 Experiments

In the following, we describe the experimental setup for quantifying the performance on basic perception tasks, advanced visual scene understanding, as well as the comparison of specialized medical VLMs and generalist VLM.

RQ1: Basic Perception Tasks on Surgical Images

To answer RQ1, we compared the performance achieved by SOTA models on natural domains to that obtained for our new dataset. Specifically, we applied our framework to the eight domains depicted in Figure 4.25. Note that the questions proposed in prior work [156] were filtered by

removing questions that cannot reasonably be applied to the surgical domain (e.g. determining whether an image is rotated). A total of 6 SOTA models identified in [156] (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, Gemini 1.5 Flash, Qwen2-VL 72B and Llama 3.2 90B) were applied by prompting them with the question along with the image.

RQ2: Advanced Frame-based Visual Scene Understanding

We answered RQ2 with three different datasets (see Table 4.9), for which we generated/adopted endoscopic vision understanding tasks of different level of complexity. Sample questions are depicted in Figure 4.27. SSG-VQA [212] builds upon the existing Cholec80 [191] dataset, containing questions about objects, their attributes, and inter-object relationships. Some of the questions are structured across three complexity levels, simple, medium, and hard (called *zero-hop*, *one-hop* and *single-and* in [212]),each demanding progressively more sophisticated visual reasoning capabilities. We randomly sampled 500 frames, questions and corresponding answers from those question-answer pairs that came with complexity ratings. We leveraged Endoscapes-Critical View of Safety (CVS) [121] which is annotated with ratings for the three criteria for Critical View of Safety (CVS) for a clinically relevant highly complex task. Specifically, 1,000 questions frames were randomly sampled and converted into questions for VQA benchmarking. While the focus of this paper is on surgical applications, we expanded its scope to general endoscopic vision due to the sparsity of benchmarks in the field. Kvasir-VQA [55] contains questions on endoscopic images, encompassing both gastroscopic and colonoscopic examinations. Each image in the dataset is annotated with answers to six different types of questions, providing comprehensive coverage of clinical visual understanding tasks. We randomly selected 2,000 frames for benchmarking, while removing questions with multiple possible correct answers.

For VQA benchmarking, we expanded the set of SOTA models used to solve basic perception tasks in order to investigate scaling and reasoning behavior more deeply. Specifically, we added the 2B and 7B variants of Qwen2 VL [206], which share an identical 675M-parameter vision encoder with the 72B variant. We also included QVQ 72B, an early open-weight reasoning VLM built on Qwen2 VL 72B, to assess the impact of chain-of-thought reasoning capabilities.

We maintained consistent conditions across all evaluations by setting temperature to 0, as is customary in VLM benchmarking frameworks. For advanced tasks, the list of possible answers was appended to the prompt. All responses underwent basic post-processing answer cleaning (e.g. removal of thinking tokens) to ensure standardized evaluation.

RQ3: Comparison of Specialized Medical VLMs versus Generalist Models

In RQ3, we explicitly compared the performance of medically fine-tuned VLMs and generalist VLMs for complex tasks by choosing pairs of models with matched size and architecture. Specifically, we focused on the latest model family Mini-InternVL [24], which offers both generalist and medically tuned versions across three scales (1B, 2B, and 4B) for a direct comparison on surgical tasks.

4.4.4 Results

RQ1: Basic Perception Tasks on Surgical Images

Endoscopic vision data typically occupies a distinct encoding space compared to natural images [58], as demonstrated with the PCA visualization of the Qwen2 VL 7B vision encoder) (Figure 4.25a).

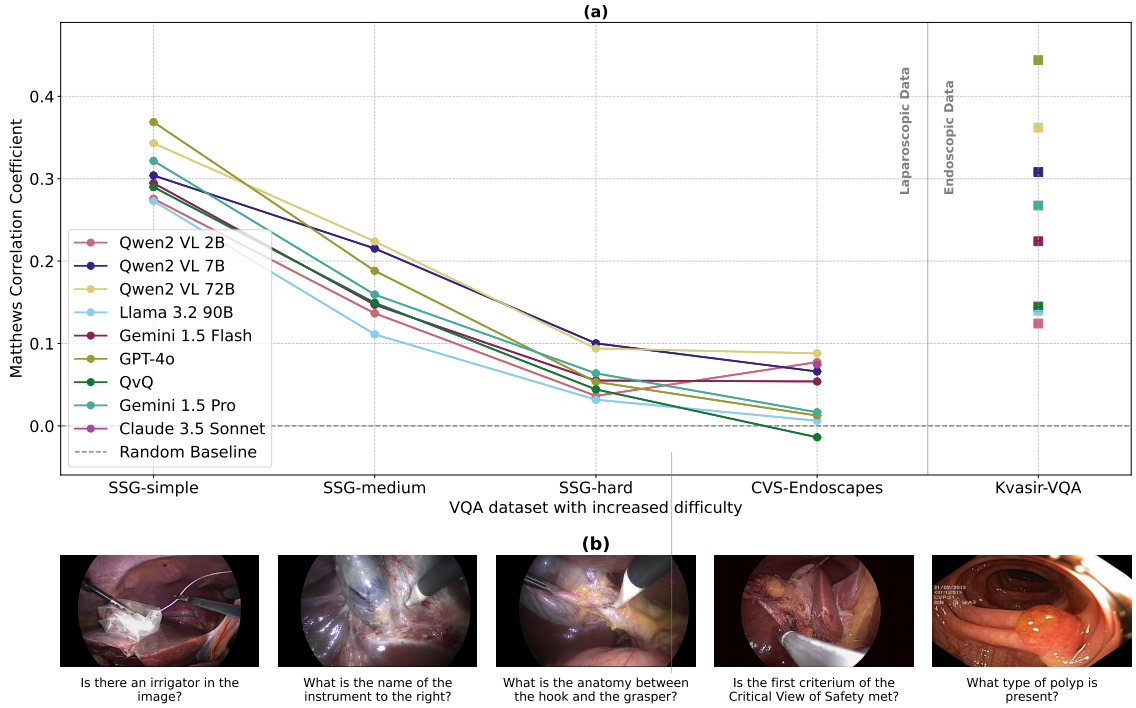


Figure 4.27: **(a) Matthews Correlation Coefficient is depicted as the function of task complexity for all advanced tasks on laparoscopic data (left) and separately for endoscopic data (right). (b) Example questions rephrased for brevity.**

Although VLMs are primarily trained on natural images, our study reveals that they process surgical images as effectively as natural images for basic perception tasks (Figure 4.25b). All models perform substantially worse than humans across all domains. The challenges encountered by VLMs when processing surgical images were found to be similar in nature to those faced when processing images from other domains (Figure 4.26).

RQ2: Advanced Frame-based Visual Scene Understanding

Model performance consistently decreased as task complexity increased across all model variants (Figure 4.27). On SSG-hard and CVS, all models achieved only marginally better results than random chance. Looking especially at the Qwen family, larger models (7B and 72B) consistently outperformed the smaller 2B variant. Notably, the QvQ model, despite being specifically designed for reasoning tasks, performed worse than both 7B and 72B models across all evaluated tasks. Testing on the Kvasir-VQA dataset confirmed these findings, demonstrating no performance improvement from reasoning-enhanced models when handling complex questions.

RQ3: Comparison of Specialized Medical VLMs versus Generalist Models

Medical foundation models consistently underperformed compared to generalist vision-language models (VLMs) across all evaluation metrics (Figure 4.28). This performance gap remained evident

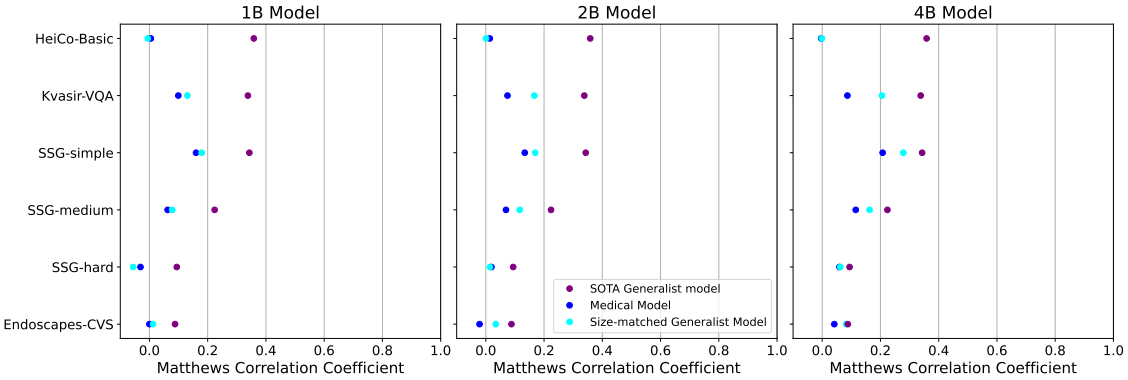


Figure 4.28: **Medically fine-tuned models fail to outperform matched generalist models on surgical questions.** For all advanced tasks, Matthews Correlation Coefficient (MCC) is depicted for the medical model, a size-matched generalist and the SOTA generalist Qwen2VL-72B.

throughout all datasets and across varying levels of task complexity. Notably, neither specialized medical models nor generalist models achieved performance comparable to Qwen2VL-72B.

4.4.5 Discussion

While our paper does not present a new method for solving a particular (class of) problem, it provides important insights to guide further research in the endoscopic vision community through the first evaluation of zero-shot question-answering VLMs.

Our findings show that generalist VLMs perform comparably in surgery as they do in other domains when addressing similar questions. Hence, rather than waiting for specialized models, researchers should be more courageous applying generalist models to their surgical data science problems. Given that their basic image understanding is already quite strong and will likely continue to improve with new base models, the key focus should be on how to inject surgical knowledge effectively. The performance gap we observed between medical and generalist models is understandable, as medical models like the medical adaption of Mini-InternVL were trained on databases with limited surgical content.

Our analysis was limited to static images, as many accessible SOTA VLMs don't yet process video input. Benchmarking questions should continue to evolve, as current versions may not fully capture the real-world complexity of surgery, and the clinical relevance of model performance may vary depending on the use case. Future research should explore surgical-specific adaptations of VLMs beyond fine-tuning, including optimized prompt design with in-context examples to better leverage surgical context. While evaluations of VQA models exist in various medical disciplines, ours is the first large-scale study focused on surgery and endoscopy. This is partly because early VLMs in surgery were limited to CLIP-based methods, such as HecVL [213], which lacked open-ended visual question-answering capabilities.

In conclusion, our study provides key insights into both the strengths and limitations of current VLMs in surgical applications. Our findings highlight the need to move beyond basic image encoding improvements and toward strategies that effectively integrate surgical expertise into large generalist models.

5 Discussion

Each of the four core contributions presented in this thesis includes a dedicated discussion of its respective results. The analysis related to labeling instructions is presented in Section 4.1.5, the findings on Quality Assurance (QA) are discussed in Section 4.2.5, the DomainBench contribution is addressed in Section 4.3.5, and the application to surgical data is examined in Section 4.4.5. In the following, we provide a unified discussion that places these individual results within a broader methodological and conceptual context. Each part of this discussion can be read independently as they cover a broad range from new training paradigms to real-world translation.

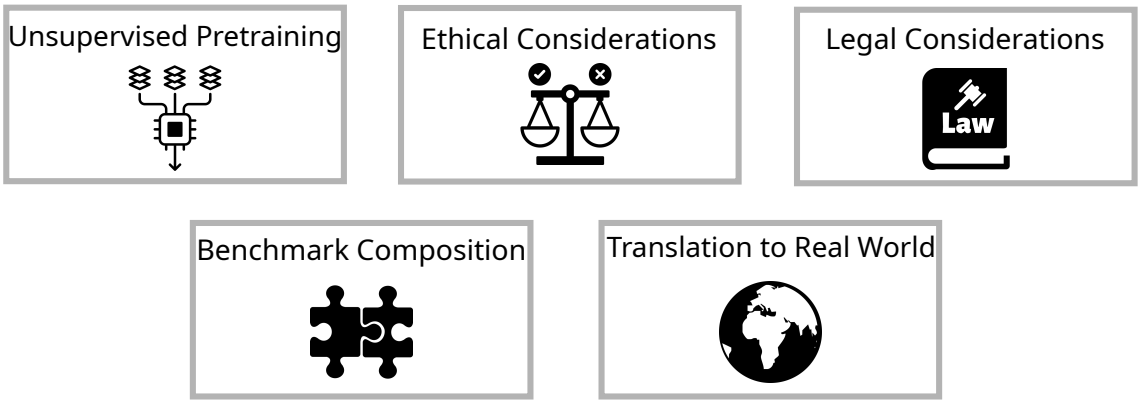


Figure 5.1: **General topics covered in the unified discussion.** Each contribution is discussed in its respective section. Here we provide a unified discussion covering (i) age of unsupervised pretraining, which no longer requires annotations; (ii) ethical considerations, which focus on annotators and their relevance for current algorithms; (iii) legal considerations, often relevant for deploying models; (iv) benchmark composition; and (v) translation to the real world, which is the goal for models.

5.1 Repositioning Human Annotation in the Age of Unsupervised Pretraining

For the past 20 years, human-annotated data served as the foundation for supervised learning, particularly in the vision and medical imaging domain, where expert judgment was required to formalize task definitions and ground truth labels [110]. Given the inherent ambiguity of medical imaging data, some researchers prefer the term reference annotation, because medical domain experts do not always agree on a joint solution for certain annotations [117]. Especially within the medical imaging domain, the ambiguity can, depending on the use case, vary drastically. For example, the inter-rater agreement for structures with clear boundaries, such as kidneys, is higher when compared to the inter-rater agreement for spreading kidney tumors [83]. Independent of the use case, annotation is associated with substantial effort, especially when domain researchers are required [83].

As a result of these and related factors, there has been a marked shift in the past five years toward unsupervised or self-supervised training, in which models learn from raw and unlabeled data without relying on explicit human supervision during pretraining. This trend is driven by multiple factors: (i) the exponential and growing volume of unlabeled data available, (ii) the empirical success of contrastive, masked, or generative objectives in Vision-Language Models (VLMs), and (iii) the scalability constraints of high-quality annotation pipelines.

Exponential data growth is occurring both within specific domains and across different fields. For general-purpose datasets, web crawling is the most predominant approach. LAION-5B [171] and The Pile [54] have demonstrated this shift by aggregating massive scale web corpora with minimal or no manual labeling, prioritizing breadth and scale over fine-grained annotations. It could be argued that the LAION dataset is at least partially annotated, because the images are provided with their captions. However, since the extraction is performed automatically and filtered via Contrastive Language-Image Pretraining (CLIP), this dataset does not contain human-annotated data in the stricter definition. In contrast to general-purpose datasets, for medical imaging data, web crawling is not a viable approach for most scenarios. Unlike general-purpose datasets, medical data primarily resides within hospitals and clinics, where access is strictly controlled due to privacy and security concerns. While specialized platforms like NIMH Data Archive exist to facilitate data sharing in healthcare, they are not directly accessible to the public and typically require formal data use agreements. This controlled access environment, while necessary for patient privacy, creates additional challenges for scaling medical model development compared to other domains. Furthermore, country-specific regulations complicate international data sharing. A prominent example is General Data Protection Regulation (GDPR), which provides rough guidelines for data sharing that need interpretation by each institution leading to broad range of interpretations even within a single institution.

As a result of the training paradigm shift, the role of human-annotated data is no longer central at the pretraining stage but is increasingly concentrated in other stages of the model development, in particular in post training, fine-tuning, alignment, and downstream evaluation. This shift changes where and how human input shapes model behavior. This makes it all the more important to understand when annotated data is still essential, and what level of clarity, consistency, and domain relevance it must provide to remain useful in modern model pipelines.

Our work has contributed substantially to the new role of human-annotated data. Labeling instruction and QA (**RQ1**) improvements enable a more fine-grained description and complete

taxonomy of recorded datasets; since edge cases are, by definition, rare and difficult to embed into a taxonomy, this enrichment yields better model generalization. Increasing adoption of models generates more real-world interactions, and hence the absolute number of edge cases rises, making our generated insights on labeling instructions, QA, and benchmarking an important contribution toward model translation. The experiments confirm the simplified message *quality beats quantity*, consistent with Molmo, the leading open-source VLM in 2024 [37], trained on only 712 k images with detailed captions exceeding two hundred words, and with s1 simple test-time scaling [132], which boosts performance by allocating extra inference compute to only one thousand reasoning-trace questions.

Furthermore, we highlight how to utilize limited resources efficiently (**RQ1-RQ2**), which is particularly relevant for academic researchers, who have less access to capital than industry researchers. The efficiency gains include the labeling instruction creation, annotation-provider selection, QA, and benchmarking of VLMs. During benchmarking, improved annotation quality and diversified tasks should surface model issues earlier, thereby accelerating the model-development cycle. In summary, we were able to narrow the gap between rapid technical progress in ML and the need for dependable, real-world solutions, especially in areas where reliability and robustness are non-negotiable, such as biomedical imaging.

5.2 Legal Considerations in Data Collection, Annotation, and Benchmarking

While the preceding discussion parts have focused on methodological, empirical, and technical aspects of data curation, the legal dimension remains an essential factor, particularly as Foundation Models (FMs) are increasingly deployed in the real-world. In the following, we outline selected legal considerations relevant to the data practices discussed in this thesis.

The collection, use, and distribution of data for training and testing FMs are increasingly constrained by legal and regulatory frameworks, particularly when the data in question involves personal, proprietary, or ethically sensitive content. In the medical domain and beyond, regulations such as the GDPR [196] in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) [47] in the United States of America impose strict conditions on the handling of patient data, including requirements for anonymization, purpose limitation, and consent. The concrete actions are often in contrast to Natural Language Processing (NLP) datasets, where surface-level metadata (e.g., names of patients, addresses, and other details) removal may suffice, medical imaging often contains visual features that can be used as indirect identifiers, rendering conventional de-identification strategies insufficient. For example, a full body MRI scan can be used to identify the face of the patient. The public distribution of (annotated) medical datasets therefore demands additional safeguards [164], including formal data use agreements, ethics approvals, and in many cases, explicit re-consent if the intended downstream use diverges from the scope of the original collection.

The use of multimodal datasets increases the complexity of data use agreements in several ways. First, the agreements need to be extended to cover all additional modalities included in the dataset, ensuring appropriate protections and usage terms for each type of data. Second, the risk of re-identification increases substantially with each added modality, as the combination of different data types can provide more opportunities for identifying individuals, which could imply necessary re-annotation of existing datasets. This heightened re-identification risk may

potentially conflict with the GDPR's principles of purpose limitation and data minimization [196], which require that data collection and processing be limited to what is necessary for specific, explicit purposes. Third, existing agreements for unimodal datasets may need to be adapted or extended, which normally triggers another round of data protection process reviews.

The legal constraints extend beyond privacy regulation. In general-purpose datasets, many of which are compiled via large-scale web crawling (e.g., LAION [171], The Pile [54]), the copyright status of included material is often unclear. Whether training models on copyrighted data constitutes lawful fair use remains an open legal question, and several ongoing lawsuits [16, 203] underscore the absence of a clear settlement in this area. To address the limitations, legal risks are further compounded in crowd-sourced annotation pipelines, where task ownership, labor conditions, and intellectual property rights associated with labels are often ambiguously defined [32, 141]. In addition, when annotation tasks involve the global distribution of sensitive data, such as medical or personalized images, to large pools of crowdworkers, the risk of unintentional data leakage increases substantially [32]. Even when formal confidentiality agreements (e.g. Non-Disclosure Agreements) are in place, varying jurisdictional guidelines and platform-level enforcement limitations complicate accountability. Institutions releasing datasets under research ethics protocols may impose non-transferable or use-case-specific usage restrictions, further limiting their appropriateness for general-purpose model training. These limitations are amplified by the opacity of data provenance in large-scale datasets, where documentation of origin, licensing conditions, and consent status is often partial or entirely absent. For example, LAION [171] specifically only provides the links to the images and not the image itself. The ambiguity is highlighted by their license terms as *"We distribute the metadata dataset (the parquet files) under the Creative Common CC-BY 4.0 license, which poses no particular restriction. The images are under their copyright."*.

This regulatory landscape creates a competitive imbalance in the global AI race. In contrast to the Western world, China has implemented minimal or no restrictions on data usage. If we assume a level playing field in terms of compute resources and technical knowledge, it can be expected that Chinese models would outperform Western models purely based on their data advantage [36]. This regulatory asymmetry raises concerns about the long-term competitiveness of Western AI development under current legal frameworks, while also highlighting the need for balanced approaches that protect individual rights without stifling innovation.

In summary, the legal concerns get more relevant with every day as FMs are increasingly integrated into everyday workflows across domains such as healthcare, education, public administration, and finance. Addressing the legal and procedural foundations of dataset construction is therefore not merely a matter of compliance but a structural prerequisite for the safe and legitimate deployment of models at scale.

While our work did not directly cover legal aspects of the data creation process, we contributed by highlighting the need for documenting and creating high-quality labeling instructions as part of the data creation process (**RQ1**), and for responsible benchmarking practices (**RQ2**). This provides practical support to help meet key compliance requirements. Specifically, our emphasis on detailed and clearly documented labeling instructions (**RQ1**) contributes directly to improving data traceability, provenance documentation, and adherence to GDPR and HIPAA requirements. Clearer instructions and documentation allow data stewards to better justify dataset usage, simplify auditing, and improve accountability. Our findings on QA (**RQ1**) additionally offer guidance for resource-limited scenarios, helping teams decide whether to prioritize better labeling instructions

or invest more into QA.

Moreover, comprehensive labeling instructions could explicitly caution annotators against unintentionally adding personally identifiable information. An additional optional flagging could also be implemented in the labeling instructions to mark cases where anonymization potentially failed (e.g., faces visible during colonoscopy or in autonomous vision scenarios). If used, such a mechanism would support downstream bias assessments and ethical compliance, particularly in multimodal datasets. Overall, bias assessments profit from more metadata about the data creation process, because they can be used as additional input factors. Finally, our recommendation of benchmarking on one's own datasets rather than relying on large-scale datasets (**RQ2**) aligns with the data minimization principle. This reduces potential legal ambiguity regarding copyright and allows researchers to spend effort on detailed annotations rather than accumulating ever-growing data volumes.

5.3 Ethical Considerations in Data Collection and Annotation

The thesis has focused mainly on the technical and empirical aspects of annotation, QA and benchmarking. In this section, we address selected ethical challenges that arise in the creation, annotation, and benchmarking of datasets. This perspective is particularly relevant in light of current industry practices, where annotations are frequently outsourced with limited oversight regarding working conditions, data protection, and task execution. The ethical implications of outsourcing affect not only the annotators themselves but also the resulting datasets and models, with consequences for both data quality and fairness among others.

Over the last four years, the annotation market has undergone seismic changes driven largely by general algorithmic advances, as highlighted in Section 2.5. Another minor impact factor is the evolving legal framework governing data collection, annotation, and benchmarking, addressed earlier in this discussion. Together, these developments redefine the requirements placed on annotators and, in turn, the employment relationship. In this part of the discussion we differentiate between expert knowledge and laymen knowledge. We define laymen knowledge as knowledge that comprises facts or explanations that a typical, non-specialist can reasonably provide without formal training. In contrast, expert knowledge consists of insights or solutions that require specialized education, extensive practice, or domain-specific experience and are therefore accessible only to a comparatively small subset of people.

For supervised computer vision scenarios, similar to the experiments conducted for RQ1, the common approach to crowdsourcing was, and continues to be, to break down the annotation task and transfer it to a large group of annotators. The annotators are provided with labeling instructions and tooling to perform the annotation. In this scenario, most annotators are considered laymen, as practically anyone can perform the required annotations. However, this renders annotators replaceable and results in many of them facing low wages and precarious employment conditions, little institutional protection or governmental oversight [209].

With the remarkable progress on general VLMs benchmarks, such as MMMU [70] in recent times, the validation focus for VLMs is shifting from broad, general surface knowledge tasks toward expert knowledge tasks [64]. This trend is best illustrated by the evolution of included math problems, even though these tasks are primarily language-based. Whereas in 2021, GSM8K [30] centered on grade-school level math problems, more recent benchmarks increasingly incorporate United States of America Mathematical Olympiad (USAMO)-level problems [10]. Combined with

the ongoing trend toward multimodality (see Section 3.3.2), this raises a fundamental question: *When is an annotator considered an expert today?* A related concern is whether traditional crowdsourcing with laymen still holds its relevance under these evolving validation demands.

For expert knowledge VLM tasks in formal knowledge domains such as physics, mathematics, and engineering, laymen are no longer the adequate choice. The complexity of such tasks demands in-domain expertise. A nuclear engineer or virologist, for instance, requires formal education spanning several years to acquire the relevant knowledge. In contrast to traditional crowdsourcing settings, annotators in these contexts are domain experts: they are not easily replaceable, can command expert-level compensation, and are unlikely to face precarious working conditions. With the rising adoption of AI across industries [1], traditional crowdsourcing should remain applicable to many industrial annotation tasks.

The global distribution of workers is a defining characteristic of traditional crowdsourcing. Given that current state-of-the-art (SOTA) VLMs are underperforming on culturally contingent tasks [149], crowdworkers may be uniquely positioned as experts for such domain tasks. However, it remains an open question if this new expert role has the potential to improve their working conditions. Exemplarily, Pouget et al. [149] showed that limiting pretraining to English image-text pairs disproportionately harms downstream recognition and reasoning tasks for culturally diverse and lower socioeconomic contexts, while standard benchmarks, such as ImageNet and COCO fail to expose the shortfall. This VLM limitation has been reported by a large number of studies [8, 87, 159]. To measure global representation, Pouget et al. [149] further propose geo-localization, defined as predicting an image's place of origin, as a task that is intrinsically sensitive to cultural context and therefore capable of revealing diversity gaps in multimodal representations. This is a promising approach, as it could be applied to other datasets that do not contain geolocation information yet. Users could extract this information from EXIF data, other metadata sources or potentially add a proxy with SOTA VLMs [75].

Following the transition from laymen knowledge to expert knowledge, the current state for benchmarking VLMs mimics the annotator requirements of the biomedical imaging domain. Whereas simple tasks, like detecting surgical instruments, may be performed by laymen [117], more complex tasks require expert medical domain knowledge [83]. The upper performance ceiling is the inter-rater agreement between the medical experts [83]. Analogously, the upper bound of achievable performance on complex VLM tasks is likely constrained by the inter-rater agreement among domain experts. Exceptions to both cases would include verifiable tasks, wherein an objective ground truth can be obtained. A common example in the biomedical domain is tissue classification. While the initial model prediction is based solely on the image, the definitive ground truth is subsequently determined through pathologist grading, thereby rendering the task objectively verifiable.

Our work has made a substantial contribution to the ethical considerations in data collection and annotation. First, with the improved labeling instructions and QA (**RQ1**), annotators are able to work on less ambiguous tasks and complete them more efficiently compared to cases where little or no instruction is provided. Given the substantial power imbalance between annotators and employers, particularly in low-cost countries, this results in a substantially improved working environment by providing a more structured and transparent task context. The heightened focus on labeling instructions and QA can be used to co-develop high-quality instructions from the start of an Machine Learning (ML) project with the annotation requester. This leads to lower turnover, faster project completion, and reduced costs. Responsible benchmarking practices

(**RQ2**) similarly promote a stronger matching between presumed model improvements and real-world application performance. Before our work, users were required to manually generate their own benchmarks and largely relied on general-purpose benchmarks to estimate model performance on their specific data. With DomainBench, we narrow this gap by providing a framework that enables users to rapidly turn their own images into a benchmark. Furthermore, the use of DomainBench necessitates direct interaction with users' own data, thereby fostering greater awareness of dataset-specific characteristics and data in general. For example, data collection is conducted at the beginning of an ML project in the upstream stage and introduced errors propagate to later project stages [168]. In summary, our work raises awareness for the importance of both data collection and annotation alike - leading to improved working conditions for annotators and more robust ML project pipelines.

5.4 Benchmarking Composition in the Future

In recent years, benchmarks have changed drastically in scope, size and iteration speed among others [64]. For an overview on VLM benchmarks, we refer the reader to Section 3.3. Here, we briefly reflect on their overall development before connecting to our own contributions. As highlighted in the Related Work Section, the number and size of available benchmarks is steadily increasing. Early efforts expanded the range of domains and questions in generalist benchmarks, aiming to strengthen the signal by representing a broader set of concepts. With progress on these general benchmarks, such as MMMU [70], expert-domain benchmarks requiring deep specialist knowledge emerged, such as the Virology Capabilities Test [60]. Benchmarking has become a recurring cycle, resembling a cat-and-mouse dynamic [64], in which each new released benchmark is rapidly solved by newer models and no longer enables model performance comparison. This is particularly relevant for closed source models, where the training data and code is not available, mostly because data is one of the major competitive advantages in the race for the best performing model. However, it still remains unclear which fine-grained concepts are embedded in the VLMs and if they can extrapolate from their training data. Another major shortcoming of the rapid release of benchmarks is their limited quality, as current metrics often fail to adequately represent the underlying problems [118], particularly in complex or real-world settings.

Recent analysis from Udandarao et al. [192] of 34 VLMs pretrained on five standard web-crawled corpora demonstrates that "zero-shot" success primarily reflects the frequency of downstream concepts in the pretraining set. Their study establishes a log-linear scaling law in which each linear uptick in zero-shot accuracy demands an exponential rise in concept-specific samples, a sample-inefficient pattern that persists after controlling for near-duplicate data points and when probed with synthetic distributions. When the same models are benchmarked on the introduced long-tail "Let it Wag!" test set, performance deteriorates markedly, indicating that current large-scale web data corpora do not capture these long-tail concepts yet. One could argue that Humanity's Last Exam [146] is the solution for the long-tail problem. Humanity's Last Exam is a multimodal benchmark designed to test VLMs at the frontier of human academic knowledge. It comprises 2,500 expert-authored questions across diverse domains, including mathematics, humanities, and natural sciences, with answers that are unambiguous, verifiable, and resistant to simple retrieval. The data was collected with a worldwide crowdsourcing campaign and competitive prize pool. SOTA models available at Humanity's Last Exam release date performed poorly, indicating a persistent gap between the model capabilities and expert humans. It should

be noted that the benchmark was crafted adversarially by removing questions from the collected question pool that existing models could already answer. The best-performing model available at the release date was o1 from OpenAI, achieving 8%. With the inclusion of "Last Exam" in the benchmark title, readers might assume that this would be the final benchmark required to surpass human-level performance. Yet the authors themselves clarify in their paper: *"HLE may be the last academic exam we need to give to models, but it is far from the last benchmark for AI."* The competition between models and benchmarks continues, as within less than five months after the release of Humanity's Last Exam, frontier models had already surpassed 25%. It appears likely that future model iterations will be trained on ever larger datasets, encompassing even fine-grained concepts which were previously overlooked.

While the previous developments aim to extend the data and included concepts for the training of SOTA VLMs, parts of the AI Safety Community have advocated for a contrasting strategy that aims to enhance safety by selectively excluding particular data samples from pretraining. Open source models offer transparency and broadened research access for everyone, but they remain acutely vulnerable to tampering attacks, in which adversaries can induce harmful behaviors through relatively minor weight or activation modifications [228]. Existing post-training safeguards, such as safety fine-tuning, have exhibited limited robustness to attacks, with models typically reverting after only a small number of adversarial updates. O'Brien et al. [137] introduced a scalable multi-stage filtering pipeline and trained multiple 6.9B-parameter models on corpora with systematically excluded dual-use content. Their study demonstrated substantial resistance to adversarial fine-tuning, persisting for up to 10,000 steps and 300M tokens of biothreat-related text. These filtered models outperformed post-hoc interventions by more than an order of magnitude, while maintaining performance on unrelated capabilities. Lastly, they reveal that models can still exploit harmful knowledge through retrieval or search tools. This approach might not only prove useful for open source models, but could be applied to closed source models as well, as jailbreaking [163] remains a major issue.

In summary, we can expect that future benchmarks will assess more in-depth capabilities, continually expand their scope, and simultaneously reduce dangerous or otherwise harmful content in the training data. This section has focused so far on the composition of benchmarks. In the next part of the discussion we will address translation of research findings to real-world data.

The improved labeling instructions and QA (**RQ1**) substantially support the generation of new benchmarks when they contain human annotations. For benchmarks without human annotations, extending the documentation of data processing and selection allows users to interact with the data in a more transparent and meaningful way. In such cases, benchmark creators should publish the generated code to support reproducibility. Our proposed concept for user-centric benchmarking with DomainBench (**RQ2**) has been at the forefront of shifting benchmarking toward user requirements. The user-specific benchmarking approach has since been adopted by HuggingFace for Large Language Models (LLMs). Shashidhar et al. [175] present an open-source framework that generates reliable, domain-tailored benchmarks automatically from user-provided documents. They further demonstrate replication of established subsets such as MMLU with minimal source text and low cost while preserving model ranking fidelity. Combined with the trend toward live benchmarks, it is reasonable to anticipate a future characterized by continuously expanding and evolving benchmark resources.

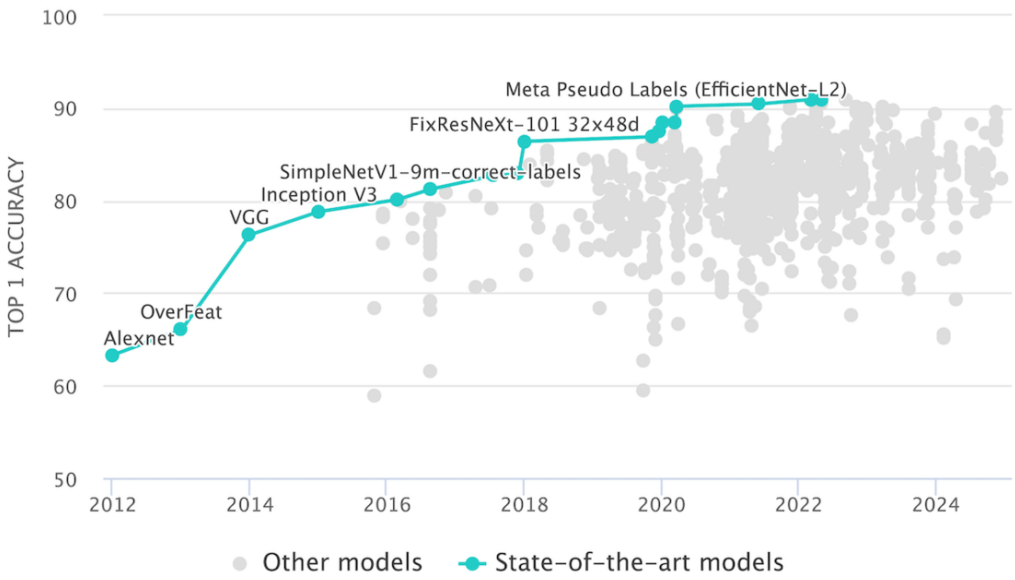


Figure 5.2: **Progress on ImageNet from 2012 until 2025.** ImageNet remains unarguably a leading benchmark for pattern recognition, as model developers incrementally compete over the best performance. Each point represents a different ML model evaluated on ImageNet with the Top 1 Accuracy. Despite its popularity, progress on ImageNet does not improve translation to real-world tasks, as displayed by Tuggener et al. [188] and Fang et al. [48]. Data and illustration from Papers with Code (license: CC BY-SA).

5.5 Translation of Research Findings to Real-World Data

We demonstrated the importance of labeling instructions and rigorous benchmarking across a diverse range of domains. From a macro perspective, the machine learning community operates on the premise that progress is measured through performance on such fixed benchmarks, where everyone can participate. Researchers and practitioners submit models, compete for the best results, and ideally, over time, solutions improve until the task is considered "solved" and ready for real-world deployment [118]. Improvements can occur either erratically or steadily, depending on the submitted solutions and task at hand. An excellent example of erratic progress, AlphaFold 2 [84] performs predictions of protein structures and participated in the 14th iteration of the Critical Assessment of Structure Prediction. It achieved a global distance test score, which measures the similarity between two protein structures with known amino acid correspondences, above 90 (max: 100). In contrast, the second place scored 68.5, and the protein folding community widely regarded the problem as effectively solved by AlphaFold.

Achieving a SOTA performance has become a common prerequisite for publication in major computer vision venues, leading to a large number of papers claiming incremental and steady gains [181]. Despite this increase of reported improvements, real-world adoption of these models remains limited and far below what the publication volume would imply. This begs the question: *Does the underlying data that researchers benchmark on represent the real world?*

Consider ImageNet [94], arguably one of the most influential benchmarks in computer vision with over 85,000 citations as of today. Initiated by Fei-Fei Li in 2006, ImageNet was constructed after

the structure of WordNet [127] and the cognitive load that humans can recognize. Built upon over 20,000 WordNet noun categories, ImageNet consists of over 14 million human-annotated images spanning more than 20,000 object categories. The WordNet categories are connected by means of conceptual-semantic and lexical relations.

Building on that question, Fang et al. [48] investigate this question by evaluating 19 ImageNet pre-trained models with accuracies between 57% and 83% on six domain image classification datasets, such as wildlife camera traps [12] and satellites [68]. Newer ImageNet-motivated architectures displayed little to no progress after the VGG architecture for four out of six downstream datasets with a fitted slope relating the ImageNet accuracy to task accuracy below 0.05. Furthermore on several datasets, task-specific improvements such as extra training data or data augmentations result in higher performance gains than updating to a more recent architecture. While their experiments include Transformer-based architectures, Tuggener et al. [188] observe similar findings on 500 CNN-based models and eight domain datasets.

A growing body of work explicitly targets performance in the wild-i.e. on data collected under the unconstrained, messy real-world conditions that represent the deployment stage. Such data contains shifts based on temporal, domain, and geographic aspects that are not included in benchmarks under regular circumstances. As a representative example, Koh et al. [92] provide ten datasets drawn from settings like pathology, camera-traps for wildlife monitoring and continuously sampled satellite imaging. Each dataset is partitioned into in-distribution and out-of-distribution splits, revealing that models trained on ImageNet degrade with a loss in performance between 20 and 50 absolute points. DomainBed [61] and the Visual Task Adaptation Benchmark [218] report similar findings.

While the in the wild work represents an empirical approach, it could be argued that the out-of-distribution (ood) body of work is rooted in statistical learning theory. It formalizes generalization under explicit distribution shifts, such as label, covariate, or concept shift, and focuses more on specific algorithms. The shifts can be defined as:

$$\underbrace{P_{\text{train}}(X) \neq P_{\text{test}}(X)}_{\text{covariate shift}} \tag{5.1}$$

$$\underbrace{P_{\text{train}}(Y) \neq P_{\text{test}}(Y)}_{\text{label shift}} \tag{5.2}$$

$$\underbrace{P_{\text{train}}(Y \mid X) \neq P_{\text{test}}(Y \mid X)}_{\text{concept shift}}. \tag{5.3}$$

where $P(X, Y)$ denotes the joint distribution, with X representing the model input and Y the corresponding model output. For a comprehensive overview about ood, we direct the reader to Yang et al. [207]. Aside from a comprehensive overview, they even provide a clear delineation from closely related tasks such as open-set recognition, anomaly detection, and outlier detection.

We observe that improvements on academic benchmarks often fail to translate into real-world performance, since many reported gains arise from benchmark-specific over-fitting. The community has attempted to mitigate this limitation by expanding benchmark diversity and scale. Compared to other domains, this issue is especially present in the medical domain, where data is sparse [210], sensor modalities are more diverse, and interpretation is more challenging compared to other domains [83].

Our work contributes to the translation of research findings to the real world in several ways. We showed that the choice of annotation company can alter the median Dice Similarity Coefficient (DSC) score by more than twenty absolute points. This discrepancy is troubling in light of the results reported by Christodoulou et al. [29], who found that the difference in mean DSC between the first- and second-ranked methods at the largest conference for biomedical imaging was only one absolute point, implying that typical claims of outperformance probably do not hold, especially for small sample sizes like in medical imaging [29]. This discrepancy may represent only one of many factors that currently hinder the translation of AI into healthcare settings [17]. We recommend that researchers and practitioners alike report as many details about the data pipeline as possible. While it may not mitigate negative effects, it would at least increase transparency and reproducibility of their findings. Compared to other domains, safety-critical domains such as medical imaging require additional certification steps prior to deployment. This includes clinical studies and further regulatory procedures necessary to obtain approval as a certified medical product. Consequently, while current standards are improving, there is still a long way to go. The limited translation of research with 1250 Food and Drug Administration (FDA)-licensed medical products to date in August 2025 [66] into practice underscores this gap. We therefore recommend publishing fewer but more thorough studies, emphasizing quality over quantity.

The improved labeling instructions and QA (**RQ1**) enlarge and harmonize the dataset taxonomy through expanded class definitions, illustrative edge cases, and potential feedback. This improved hierarchy is expected to surface edge-case conflicts, ambiguous decision boundaries, and inconsistencies in the instructions before a model is trained, preventing silent failure when deployment reaches safety-critical environments, such as biomedical imaging. A parallel effect arises from the benchmarking advances addressed in **RQ2**. The diversified and domain specific tasks surface model failures earlier in the model development cycle, when corrections and retraining remain comparatively inexpensive.

The added detail to edge-case specification shifts a subset of instances that were formerly classified as in the wild, that is, out-of-distribution, into in-distribution, thereby strengthening generalization. Additionally, the added detail from labeling instructions through to benchmarking produces a clearer definition of the underlying concept. For both label shifts and concept shifts, the added detail should enable earlier identification of emerging shifts, because ambiguous cases have been reduced. It could even be argued that the amount of falsely reported concept shifts will be reduced, because cases formerly not covered by the concept are now clearly defined and no longer annotated based on the individual annotator.

Real-world data comprise observations captured during the ordinary course of daily activities, whether actively supplied by end-users or passively recorded by systems without direct user intervention. Within this context, DomainBench (**RQ2**) stands out as our most direct contribution to translation: it equips users with a framework for evaluating models on their own domain data, verifying alignment between model capabilities and their own individual task requirements, and proceeding to practical adoption once the resulting metrics satisfy application-specific thresholds.

5.6 Future Directions

We have introduced methods that substantially increased the impact and quality of validation for biomedical imaging algorithms and VLMs alike; yet considerable room for further improvement remains, and several open research questions are still unanswered. In the following, we outline four key directions for future research that build on our work.

Automation and Human-AI Collaboration

An open question is the impact of increasing automation on annotation performance. For all our experiments on labeling instructions and QA, annotators used minimal to no automation during their work. There is considerable evidence in the Human-Computer Interaction literature [179, 130] about automation bias, where participants manifested errors of omission, overlooking events that the automation failed to flag explicitly, and errors of commission, adopting the automation's recommendations even when those suggestions were erroneous. However, this has not been applied to varying labeling instructions in the ML community yet. Future work should systematically study how automation affects annotation quality, instructions, and annotator.

Toward Multimodal Data and Industry-Wide Impact

Another promising research direction is the extension of our experiments to 3D, other 2D non-RGB imaging data and ultimately multimodal data. For the rebuttal to *Quality Assured: Rethinking Annotation Strategies in Imaging AI* we conducted a survey among automotive and tech companies to estimate the potential impact of our findings beyond biomedical imaging. All participants indicated that the obtained results on biomedical imaging data hold the potential to produce practice-changing impact with respect to annotation processes in the automotive and technology industry. Conducting systematic experiments in these data domains could validate these initial results and unearth industry-specific insights.

Toward Statistically Grounded Evaluation in ML

Our work demonstrated that individual upstream components of the model development pipeline, such as labeling instructions, can have a substantial impact on the resulting reference annotations. Accordingly, the analysis of model performance should incorporate formal statistical evaluation to assess the robustness and reliability of observed effects. Yet, the broader ML community often neglects to report statistical measures. For instance, Anthropic, a leading VLM company, recently advocated for the routine inclusion of error bars in performance reporting [126]. This shortcoming is present in biomedical imaging research at the field's largest conference as well. Only a single paper reported confidence intervals in 2023 [29]. Consequently, future work should systematically vary individual components of the data pipeline and quantify their effects on downstream task performance using appropriate statistical methods.

From DomainBench to Multi-Agent and Multimodal Benchmarks

Lastly, this thesis proposed DomainBench, a task augmentation framework to generate over 20 diverse perception tasks from individual user-provided images for VLMs and applied to the general vision and the surgical imaging domain. The tasks were designed to be consistent across

domains to enable performance. A conceptually similar strategy has recently been developed in the NLPs community for basic NLP tasks such as ingestion, summarization, and question generation. Future work may extend this approach to more complex, agentic tasks involving both textual and visual inputs.

Taken together, these research directions highlight that sustainable progress in ML depends on more than incremental improvements in model architectures. By investigating automation effects on annotation, extending experiments across modalities and industries, embedding statistical rigor into evaluation, and evolving DomainBench toward interactive multimodal benchmarks, future research can move closer to the overarching goal of this thesis: aligning methodological innovation with real-world translation.

6 Summary of Contributions and Conclusion

This thesis examined several often overlooked aspects of the model development pipeline. Subsequently, several advances for the upstream stage before the model training and the downstream stage after the model training were proposed and evaluated. Combined, this thesis resulted in

- **8 main-author publications** (1 full paper in a peer-reviewed journal; 1 full paper at a peer-reviewed conference; 4 short papers at conference workshops; 1 preprint submitted to a peer-reviewed journal; 1 preprint submitted to a peer-reviewed conference),
- **7 co-author publications** (3 in peer-reviewed journals; 2 submitted to peer-reviewed conference proceedings),
- **10 invited talks at national and international events**,
- **6 awards (combined endowment over EUR 12,000)**, and
- **1 co-organized challenge at the internationally recognized conference MICCAI**

as detailed in the following Section 6.1. The main-author papers (excluding the short papers) are presented separately in Section 4. In the following I give an overview of my publications in Section 6.2 and end with a conclusion to the thesis in Section 6.3.

6.1 Summary of Contributions

Research Question 1

How can data annotation processes be optimized for reliability and efficiency?

Contribution 1: Large-scale empirical study of labeling instructions in biomedical imaging As part of this thesis, the first large-scale study on labeling instructions for biomedical imaging was conducted, involving more than 850 annotators, several labeling instructions and approximately

10,000 images. It was found that including visual examples in labeling instructions substantially improved annotation quality without adding overhead, whereas extending textual descriptions did not yield improvements. Professional annotation companies consistently produced higher quality annotations across all scenarios compared to the then-leading crowdsourcing platform Amazon Mechanical Turk (MTurk). Finally, the variability in annotations introduced by labeling instructions has been shown to exceed the performance gap between Machine Learning (ML) models.

Contribution 2: Large-scale survey on professional annotator involvement in instruction design It also observed that more than 75% of biomedical imaging competitions provided no labeling instructions. This prompted us to examine the role of annotators in data generation. The largest professional annotator survey to date, with over 300 participants, revealed that annotators are rarely involved in the creation of labeling instructions. Furthermore, labeling instructions were mentioned as the leading source of daily annotation problems.

Contribution 1 and 2 resulted in two short papers presented at *Conference on Neural Information Processing Systems (NeurIPS)* [153, 154], including one oral presentation, through which we collected feedback from the research community. Our subsequent full paper appeared in *Nature Machine Intelligence* [155]. The medical imaging community received our results positively, and following a comprehensive review by the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) Special Interest Group for Biomedical Image Analysis Challenges [119], future challenges are now required to meet stricter standards, including the mandatory publication of labeling instructions. In recognition of this lasting contribution to the field, we were awarded the *Anton Fink Science Prize for Artificial Intelligence (AI)* [95], endowed with EUR 10,000.

Contribution 3: Large-scale empirical study of internal Quality Assurance (QA) in biomedical imaging Following the findings from the annotation stage, we shifted our focus to the internal QA stage, which typically follows annotation. The setup of this study extended the previous design by incorporating an additional QA step. Our work has shown that annotation companies deliver both higher quality and higher quantity at the same time. Earlier studies had not accounted for the monetary cost differences between annotation providers. Most importantly, the internal QA implemented by professional annotation companies yielded only marginal improvements compared to the initial annotations. We also found that research on professional annotation companies has been relatively under-reported, despite the fact that they account for approximately 20% of exhibitors at major computer vision conferences.

Contribution 4: Revelation of relevant image characteristics for QA While internal QA only marginally improved the annotation quality for all images, we have identified specific image characteristics that benefit from additional QA. These characteristics were general characteristics and not limited to specific domains, enhancing the applicability beyond biomedical imaging data.

Contributions 3 and 4 resulted in a publication at European Conference on Computer Vision (ECCV) [157]. For my overall contributions to research on annotation procedures, I was invited to give talks at the Alexander von Humboldt Foundation visit at German Cancer Research Center (DKFZ), the Lindau Nobel Laureate Delegation visit at DKFZ, the Pelagic Imaging Consortium Workshop on PIC meets HI in Hamburg, and the In the Picture: Medical Imaging Datasets Workshop in Copenhagen, as well as to present a poster at the Strategic Advisory Board

Helmholtz Health Meeting in Berlin. My work was further recognized with the 1st Place Flash Talk Award at the ELLIS and AIHealth AI inScide Out Conference and the Falling Walls Science Breakthrough of the Year 2023 Finalist title.

Research Question 2

How can annotated data be most effectively leveraged for robust and domain-aware benchmarking?

Contribution 5: Identification of weaknesses in VLM benchmarking A recurring problem is that existing Vision-Language Model (VLM) benchmarks do not reflect the needs of domain users, complicating model selection. By generating a broad set of tasks across different domains, we found that VLM families exhibit characteristic weaknesses both across families and within individual families. We demonstrated how users can leverage this information for their benchmarking to select models that perform with higher accuracy on their data, which general benchmarks might not reveal.

Contribution 6: Framework for resource-efficient creation of user-centric VLM benchmarks To address this problem, we developed a framework for the resource-efficient creation of user-centric VLM benchmarks that generates multiple tasks from a single task, thereby substantially reducing the gap between domain users and their benchmarking practices. We showed that with our method, additional benchmarking information can be extracted from any existing instance segmentation dataset. To facilitate adoption of the framework, we released seven new datasets covering a wide range of domains from kitchen environments to animal scenes. To complement the generation process, we introduced a new metric, Accuracy%(t), that accounts for shared base images across different tasks.

Contribution 5 and 6 were published as two short papers at the International Conference on Learning Representations (ICLR) in 2025 and the full paper is available as a preprint [122] and currently under review at NeurIPS.

Contribution 7: First comprehensive assessment of VLMs in surgical imaging We carried out the first broad evaluation of VLM on surgical data, benchmarking 15 SOTA models across more than 167,000 question-answer pairs from multiple endoscopic datasets. The study revealed that VLMs can handle elementary surgical perception tasks, such as instrument counting and spatial localization, with accuracy comparable to natural image domains. Yet performance declines considerably once questions go beyond the elementary level and require medically informed reasoning, underlining the heightened complexity of surgical questions compared to standard benchmarks.

Contribution 8: Critical comparison of medical versus generalist VLMs Through size-controlled experiments, we systematically compared medically fine-tuned VLMs to their general-purpose counterparts. Across all tested datasets and difficulty levels, specialized medical models lagged behind generalist ones, including models that were explicitly fine-tuned with biomedical data. The findings indicate that current shortcomings are rooted less in visual representation than in the absence of robust domain knowledge. This calls for future work on knowledge integration strategies that extend beyond conventional finetuning.

Contributions 7 and 8 were published as a preprint [122] and are under peer review at Medical Image Analysis. The accompanying HeiCo-VQA-Base dataset, comprising 24,252 images and 167,384 questions with human reference annotations, now serves as a foundation for rigorous benchmarking of VLMs for endoscopic and surgical use cases.

6.2 Own Publications

This section presents all manuscripts to which I have contributed either as author or co-author. The first part comprises papers where I am the main author or core contributor, including peer-reviewed journal articles, peer reviewed conference articles, as well as preprints. Short papers are marked accordingly. The second part contains the remaining publications, where I contributed as a co-author. For each entry, the main author(s) are underlined and my own name highlighted in boldface. For preprints, the publication venue to which the manuscript has been submitted is specified.

Publications as Core Contributor

1. **Tim Rädsch**, et al. *Labelling instructions matter in biomedical image analysis*. **Nature Machine Intelligence**, 5:273–283, 2023.
2. **Tim Rädsch**, et al. *Quality Assured: Rethinking Annotation Strategies in Imaging AI*. In **ECCV**, 2024.
3. **Tim Rädsch**[1], Leon Mayer[1], et al. *DomainBench: A Framework for Scalable VLM Perception Benchmarking on Your Own Images*. arXiv preprint arXiv:2502.15563, 2025. [**submitted to NeurIPS**]
4. Leon Mayer[1], **Tim Rädsch**[1], et al. *Challenging Vision-Language Models with Surgical Data: A New Dataset and Broad Benchmarking Study*. arXiv preprint arXiv:2506.06232, 2025. [**submitted to Medical Image Analysis**]
5. **Tim Rädsch**, et al. *Labeling instructions matter in biomedical image analysis*. **Medical Imaging meets NeurIPS** (short paper), 2022.
6. **Tim Rädsch**, et al. *Labeling instructions matter in biomedical image analysis. An annotator-centric perspective*. **Human-Centered AI Workshop at NeurIPS** (short paper), 2022.
7. **Tim Rädsch**[1], Leon Mayer[1], et al. *Domain-Specific Benchmarking of Vision-Language Models: A Task Augmentation Framework Using Metadata*. In 2nd Workshop on Navigating and Addressing Data Problems for Foundation Models (DATA-FM), **International Conference on Learning Representations (ICLR**) (short paper), 2025.
8. **Tim Rädsch**[1], Leon Mayer[1], et al. *Bridging vision language model (VLM) evaluation gaps with a framework for scalable and cost-effective benchmark generation*. In Foundation Models in the Wild Workshop), **International Conference on Learning Representations (ICLR**) (short paper), 2025.

Publications as Co-Author

1. Amelia Jiménez-Sánchez, ..., **Tim Rädsch**, et al. *In the picture: Medical imaging datasets, artifacts, and their living review*. In Proceedings of the ACM **Conference on Fairness, Accountability, and Transparency (FAccT)**, pp. 511–531, 2025.
2. Lena Maier-Hein, Annika Reinke, ..., **Tim Rädsch**, et al. *Metrics reloaded: recommendations for image analysis validation*. **Nature Methods**, 21(2):195–212, 2024.

[1]Shared core contribution.

3. Annika Reinke, Minu D. Tizabi, ..., **Tim Rädsch**, et al. *Understanding metric-related pitfalls in image analysis validation*. **Nature Methods**, 21(2):182–194, 2024.
4. Nicholas Heller, Fabian Isensee, **Tim Rädsch**, et al. *Kidney and Kidney Tumor Segmentation: MICCAI 2023 Challenge, KiTS 2023*. In **MICCAI Workshops**, Lecture Notes in Computer Science (LNCS) 14540. Springer, 2024.
5. Matthias Eisenmann, ..., **Tim Rädsch**, et al. *Why is the winner the best?*. In Proceedings of the IEEE/CVF **Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 19955–19966, 2023.
6. Annika Reinke, ..., **Tim Rädsch**, et al. *Metrics reloaded: a new recommendation framework for biomedical image analysis validation*. In **Medical Imaging with Deep Learning (MIDL**) (short paper), 2022.

6.3 Conclusion

Advances in ML hold the potential to transform science and society, yet this requires effective translation of research findings into real-world impact. Despite a rapidly growing number of publications that claim improvements through new state-of-the-art (SOTA) methods, translation remains minimal. As highlighted in this thesis, the central challenge lies in the complexity of the model development pipeline, where both upstream stage tasks (problem definition, data selection, labeling instructions, quality assurance) and downstream stage tasks (benchmarking, validation, robustness analysis, monitoring) introduce design decisions that profoundly shape the outcome and ultimately the translation of a trained model.

This thesis has shown that even individual design choices, in these stages can outweigh the performance gains of claimed novel algorithms. Most prominently, we demonstrated that labeling instructions exert a substantially larger influence on annotation quality and downstream model accuracy than the incremental improvements usually reported in methodological work [29]. Our studies revealed the important role of visual examples in instructions and the limited effectiveness of internal QA practices, providing practical guidance for data annotation.

On the downstream side, we introduced DomainBench, a framework for constructing resource-efficient, user-centric VLM benchmarks through task augmentation. By generating diverse tasks from a single instance-segmented image, the framework allows domain researchers to evaluate models more effectively. Applying DomainBench to surgical data, we conducted the first large-scale audit of VLMs in endoscopy, showing that while current models succeed on basic perception, they fail on medically informed reasoning tasks. These findings emphasize the need for strategies that integrate domain knowledge into generalist models.

In sum, the thesis highlights how upstream and downstream refinements, often overlooked in favor of claimed architectural novelty, can yield advances with far greater translational relevance. By addressing labeling instructions, QA, benchmarking, and surgical validation, we contribute conceptual insights and practical tools that bring the whole field closer to bridging the gap between methodological progress and real-world application for natural and medical imaging alike.

Returning to the observation by Andrej Karpathy that progress in ML relies as much or even more on data and evaluation as on modeling itself, this thesis has provided concrete evidence

and methodological tools to substantiate that claim. By examining labeling instructions and QA upstream, and by introducing DomainBench and conducting domain-specific validation downstream, we have shown how deliberate attention to these tasks can yield advances with greater translational impact than many algorithmic improvements. In this sense, the work presented here not only reiterates but extends Karpathy's perspective, demonstrating how careful design of data and evaluation practices can decisively shape the trajectory from research to real-world application.

Bibliography

[1] Ajay Agrawal, Joshua S. Gans, and Avi Goldfarb. Artificial intelligence adoption and system-wide change. *Journal of Economics & Management Strategy*, 33(2):327–337, 2024. ISSN 1530-9134. doi: 10.1111/jems.12521. 107

[2] Mercor AI. Mercor. https://mercor.com/, August 2025. 29

[3] Mercor AI. Software Engineer - US - Mercor Jobs. mercor.com/jobs/list_AAABmHHTW6qCYjpceqhLtYJD, August 2025. 29

[4] UK AI Security Institute. Inspect AI: Framework for Large Language Model Evaluations, May 2024. 40

[5] Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. CLIP the Bias: How Useful is Balancing Data in Multimodal Learning?, March 2024. 25, 31

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A Visual Language Model for Few-Shot Learning. *International Conference on Neural Information Processing Systems*, April 2022. 22, 23, 24

[7] Reem Aleithan, Haoran Xue, Mohammad Mahdi Mohajer, Elijah Nnorom, Gias Uddin, and Song Wang. SWE-Bench+: Enhanced Coding Benchmark for LLMs, October 2024. 39, 40

[8] Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. See It from My Perspective: Diagnosing the Western Cultural Bias of Large Vision-Language Models in Image Understanding, June 2024. 107

[9] Ariful Azad and Afeefa Banu. Publication Trends in Artificial Intelligence Conferences: The Rise of Super Prolific Authors, November 2024. 1

[10] Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on Uncontaminated Math Competitions, May 2025. 106

[11] Manuel Beck. A Brief History of Time, Human Potential, and the Development of AI. In Manuel Beck, editor, *Artificial Intelligence in Sales: AI Solutions for Every Stage of the Sales Cycle: Use Cases, Tools, and Implementation*, pages 3–13. Springer Fachmedien, Wiesbaden, 2025. ISBN 978-3-658-48033-2. doi: 10.1007/978-3-658-48033-2_1. 28

[12] Sara Beery, Grant van Horn, and Pietro Perona. Recognition in Terra Incognita, July 2018. 111

[13] Simone Bianco, Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. An interactive tool for manual, semi-automatic and automatic video annotation. *Computer Vision and Image Understanding*, 131:88–99, February 2015. ISSN 1077-3142. doi: 10.1016/j.cviu.2014.06.015. 30

[14] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. 23, 74

[15] Jonathan Bragg, Mausam, and Daniel S. Weld. Sprout: Crowd-Powered Task Design for Crowdsourcing. *User Interface Software and Technology*, 2018. doi: 10.1145/3242587.3242598. 35, 46, 63

[16] Blake Brittan. Anthropic wins key US ruling on AI training in authors' copyright lawsuit | Reuters. https://www.reuters.com/legal/litigation/anthropic-wins-key-ruling-ai-authors-copyright-lawsuit-2025-06-24/, June 2025. 105

[17] Valerie K. Bürger, Julia Amann, Cathrine K. T. Bui, Jana Fehr, and Vince I. Madai. The unmet promise of trustworthy AI in healthcare: Why we fail at clinical translation. *Frontiers in Digital Health*, 6, April 2024. ISSN 2673-253X. doi: 10.3389/fdgth.2024.1279629. 112

[18] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 28

[19] Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, Wenxuan Zhang, Lifu Huang, Muhao Chen, Lei Hou, Qianru Sun, Xingjun Ma, Zuxuan Wu, Min-Yen Kan, David Lo, Qi Zhang, Heng Ji, Jing Jiang, Juanzi Li, Aixin Sun, Xuanjing Huang, Tat-Seng Chua, and Yu-Gang Jiang. Toward Generalizable Evaluation in the LLM Era: A Survey Beyond Benchmarks, April 2025. 29

[20] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2334–2346, New York, NY, USA, May 2017. Association for Computing Machinery. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3026044. 35

[21] Eric Z. Chen and Hongzhe Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics (Oxford, England)*, 32(17):2611–2617, September 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw308. 49

[22] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. AlpaGasus: Training A Better Alpaca with Fewer Data, November 2023. 74

[23] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code, July 2021. 36

[24] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks, January 2024. 99

[25] Veronika Cheplygina, Adria Perez-Rovira, Wieying Kuo, Harm A. W. M. Tiddens, and Marleen de Bruijne. Early Experiences with Crowdsourcing Airway Annotations in Chest CT. *arXiv:1706.02055 [cs]*, 10008: 209–218, 2016. doi: 10.1007/978-3-319-46976-8_22. 46, 63

[26] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, March 2024. 77

[27] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying Vision-and-Language Tasks via Text Generation, May 2021. 24

[28] Hyeong Kyu Choi, Maxim Khanov, Hongxin Wei, and Yixuan Li. How Contaminated Is Your Benchmark? Quantifying Dataset Leakage in Large Language Models with Kernel Divergence, May 2025. 37

[29] Evangelia Christodoulou, Annika Reinke, Rola Houhou, Piotr Kalinowski, Selen Erkan, Carole H. Sudre, Ninon Burgos, Sofiène Boutaj, Sophie Loizillon, Maëlys Solal, Nicola Rieke, Veronika Cheplygina, Michela Antonelli, Leon D. Mayer, Minu D. Tizabi, M. Jorge Cardoso, Amber Simpson, Paul F. Jäger, Annette Kopp-Schneider, Gaël Varoquaux, Olivier Colliot, and Lena Maier-Hein. Confidence intervals uncovered: Are we ready for real-world medical imaging AI?, September 2024. 1, 4, 112, 113, 120

[30] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021. 36, 106

[31] Cognilytica. Data Preparation & Labeling for AI 2020. https://www.cognilytica.com/2020/01/31/data-preparation-labeling-for-ai-2020/, 2020. 46, 71

[32] Perrine Créquit, Ghizlène Mansouri, Mehdi Benchoufi, Alexandre Vivot, and Philippe Ravaud. Mapping of Crowdsourcing in Health: Systematic Review. *Journal of Medical Internet Research*, 20(5):187, 2018. doi: 10.2196/jmir.9330. 27, 56, 105

[33] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques and Assurance Actions. *ACM Computing Surveys*, 51(1):1–40, January 2019. ISSN 0360-0300, 1557-7341. doi: 10.1145/3148148. 14, 31

[34] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates

using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28, 1979. 30, 34

[35] J. C. F. de Winter, M. Kyriakidis, D. Dodou, and R. Happee. Using CrowdFlower to Study the Relationship between Self-reported Violations and Traffic Accidents. *Procedia Manufacturing*, 3:2518–2525, January 2015. ISSN 2351-9789. doi: 10.1016/j.promfg.2015.07.514. 27

[36] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. 105

[37] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, December 2024. 87, 89, 104

[38] Jean-Stanislas Denain. Accuracy increases with estimated training compute, 2024. 40

[39] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating Data Contamination in Modern Benchmarks for Large Language Models, April 2024. 37

[40] Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. COCONut: Modernizing COCO Segmentation, April 2024. 84

[41] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation. *arXiv preprint arXiv:2112.04554*, 2021. 27

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. 35

[43] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 52, 65

[44] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 238–247, Republic and Canton of Geneva, CHE, May 2015. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-3469-3. doi: 10.1145/2736277.2741685. 27

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. 21

[46] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1013–1022, New York, NY, USA, February 2012. Association for Computing Machinery. ISBN 978-1-4503-1086-4. doi: 10.1145/2145204.2145355. 14, 35

[47] Peter F Edemekong, Pavan Annamaraju, and Micelle J Haydel. Health insurance portability and accountability act. 2018. 104

[48] Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on ImageNet transfer to real-world datasets?, January 2023. 110, 111, 154

[49] François Fleuret. *The Little Book of Deep Learning*. 2024. 22, 150

[50] Mark Forrester, Mark Forrester, and Mark Forrester. *Forrester's Pictorial Miscellany for the Family Circle*. F. & G.C. Rand, Boston, 1855. doi: 10.5962/bhl.title.21616. 28, 150

[51] Beverly Freeman, Naama Hammel, Sonia Phene, Abigail Huang, Rebecca Ackermann, Olga Kanzheleva, Miles Hutson, Caitlin Taggart, Quang Duong, and Rory Sayres. Iterative Quality Control Strategies for Expert Medical Image Labeling. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9:60–71, October 2021. 13, 56, 58

[52] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive, July 2024. 77, 96

[53] Snehalkumar "Neil" S. Gaikwad, Nalin Chhibber, Vibhor Sehgal, Alipta Ballav, Catherine Mullings, Ahmed Nasser, Angela Richmond-Fuller, Aaron Gilbee, Dilrukshi Gamage, Mark Whiting, Sharon Zhou, Sekandar Matin, Senadhipathige Niranga, Shirish Goyal, Dinesh Majeti, Preethi Srinivas, Adam Ginzberg, Kamila Mananova, Karolina Ziulkoski, Jeff Regino, Tejas Sarma, Akshansh Sinha, Abhratanu Paul, Christopher Diemert, Mahesh Murag, William Dai, Manoj Pandey, Rajan Vaish, and Michael Bernstein. Prototype Tasks: Improving Crowdsourcing Results through Rapid, Iterative Task Design. *arXiv:1707.05645 [cs]*, July 2017. 35

[54] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling, December 2020. 103, 105

[55] Sushant Gautam, Andrea M Storås, Cise Midoglu, Steven A Hicks, Vajira Thambawita, Pål Halvorsen, and Michael A Riegler. Kvasir-vqa: A text-image pair gi tract dataset. In *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, pages 3–12, 2024. 99

[56] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018. 1, 4

[57] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 2013. 84

[58] Patrick Godau, Akriti Srivastava, Tim Adler, and Lena Maier-Hein. Beyond Knowledge Silos: Task Fingerprinting for Democratization of Medical Imaging AI, December 2024. 99

[59] Patrick Godau, Piotr Kalinowski, Evangelia Christodoulou, Annika Reinke, Minu Tizabi, Luciana Ferrer, Paul Jäger, and Lena Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment. *Medical Image Analysis*, 102:103504, May 2025. ISSN 1361-8415. doi: 10.1016/j.media.2025.103504. 14

[60] Jasper Götting, Pedro Medeiros, Jon G. Sanders, Nathaniel Li, Long Phan, Karam Elabd, Lennart Justen, Dan Hendrycks, and Seth Donoughe. Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark, April 2025. 5, 12, 108

[61] Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization, July 2020. 111

[62] Dongyoon Han, Junsuk Choe, Seonghyeok Chun, John Joon Young Chung, Minsuk Chang, Sangdoo Yun, Jean Y. Song, and Seong Joon Oh. Neglected Free Lunch – Learning Image Classifiers Using Annotation Byproducts, July 2023. 35

[63] Nicolai Häni, Pravakar Roy, and Volkan Isler. MinneApple: A Benchmark Dataset for Apple Detection and Segmentation. *IEEE Robotics and Automation Letters*, 5(2):852–858, April 2020. ISSN 2377-3766. doi: 10.1109/LRA.2020.2965061. 86, 153

[64] Moritz Hardt. The emerging science of machine learning benchmarks, 2025. 1, 16, 29, 36, 37, 106, 108

[65] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. 21

[66] Center for Devices and Radiological Health. Artificial Intelligence-Enabled Medical Devices. *FDA*, Thu, 07/10/2025 - 09:48. 112

[67] Eric Heim, Tobias Roß, Alexander Seitel, Keno März, Bram Stieltjes, Matthias Eisenmann, Johannes Lebert, Jasmin Metzger, Gregor Sommer, and Alexander W. Sauter. Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging*, 5(3):034002, 2018. 32, 35, 46, 63

[68] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, July 2019. ISSN 2151-1535. doi: 10.1109/JSTARS.2019.2918242. 111

[69] Dan Hendrycks. *Introduction to AI Safety, Ethics, and Society*. Taylor & Francis, 2025. 12

[70] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021. 35, 36, 106, 108

[71] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset, November 2021. 37

[72] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022. 35

[73] Wei-Ning Hsu and Hsuan-Tien Lin. Active Learning by Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), February 2015. ISSN 2374-3468. doi: 10.1609/aaai.v29i1.9597. 30

[74] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]*, January 2018. 21

[75] Jingyuan Huang, Jen-tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. VLMs as GeoGuessr Masters: Exceptional Performance, Hidden Biases, and Privacy Risks, May 2025. 107

[76] Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI, June 2024. 37

[77] Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, February 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.359.6377.725. 32

[78] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67, Washington DC, July 2010. ACM. ISBN 978-1-4503-0222-7. doi: 10.1145/1837885.1837906. 32

[79] Lilly Irani. The cultural work of microwork. *New Media & Society*, 17(5):720–739, May 2015. ISSN 1461-4448. doi: 10.1177/1461444813511926. 72

[80] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. 31

[81] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jäger. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15009, pages 488–498. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-72113-7 978-3-031-72114-4. doi: 10.1007/978-3-031-72114-4_47. 21

[82] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.215. 39, 192

[83] Leo Joskowicz, D. Cohen, N. Caplan, and J. Sosna. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3):1391–1399, March 2019. ISSN 0938-7994, 1432-1084. doi: 10.1007/s00330-018-5695-5. 4, 5, 13, 29, 33, 103, 107, 111

[84] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. 110

[85] Kampf Kampf and Nicole Brichtova. Experiment with Gemini 2.0 Flash native image generation - Google Developers Blog. https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation/, March 2025. 1

[86] Sayash Kapoor and Arvind Narayanan. Leakage and the Reproducibility Crisis in ML-based Science, July 2022. 1

[87] Antonia Karamolegkou, Phillip Rust, Yong Cao, Ruixiang Cui, Anders Søgaard, and Daniel Hershcovich. Vision-Language Models under Cultural and Inclusive Considerations, July 2024. 107

[88] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101759. 57

[89] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4): 614–629, October 2020. ISSN 2049-8470, 2049-8489. doi: 10.1017/psrm.2020.6. 27, 46

[90] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. 20

[91] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, April 2023. 79

[92] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664. PMLR, July 2021. 111

[93] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, February 2016. 38, 192

[94] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017. ISSN 0001-0782, 1557-7317. doi: 10.1145/3065386. 18, 19, 20, 57, 110

[95] Joerg Kunz. 10.000 Euro: Anton Fink Wissenschaftspreis für Künstliche Intelligenz zum zweiten Mal vergeben. https://nachrichten.idw-online.de/2023/03/06/10-000-euro-anton-fink-wissenschaftspreis-fuer-kuenstliche-intelligenz-zum-zweiten-mal-vergeben, June 2023. 116

[96] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, July 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-020-01316-z. 12, 38, 149, 192

[97] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI Ability to Complete Long Tasks, March 2025. 37, 38, 150

[98] Thomas A. Lampert, André Stumpf, and Pierre Gançarski. An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. *IEEE Transactions on Image Processing*, 25(6): 2557–2572, June 2016. ISSN 1941-0042. doi: 10.1109/TIP.2016.2544703. 56

[99] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, volume 2126, pages 22–32, 2010. 32, 33

[100] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002. 27, 37

[101] Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. Platypus: Quick, Cheap, and Powerful Refinement of LLMs, August 2023. 74

[102] Thomas S. Lendvay, Lee White, and Timothy Kowalewski. Crowdsourcing to Assess Surgical Skill. *JAMA Surgery*, 150(11):1086–1087, November 2015. ISSN 2168-6254. doi: 10.1001/jamasurg.2015.2405. 56

[103] Chunyi Li, Xiaozhe Li, Zicheng Zhang, Yuan Tian, Ziheng Jia, Xiaohong Liu, Xiongkuo Min, Jia Wang, Haodong Duan, Kai Chen, and Guangtao Zhai. Information Density Principle for MLLM Benchmarks, March 2025. 5

[104] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, September 2003. 27

[105] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, June 2023. 24

[106] Shijie Lian, Ziyi Zhang, Hua Li, Wenjie Li, Laurence Tianruo Yang, Sam Kwong, and Runmin Cong. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29545–29559. PMLR, July 2024. 86, 153

[107] Hsuan Wei Liao, Christopher Klugmann, Daniel Kondermann, and Rafid Mahmood. Minority Reports: Balancing Cost and Quality in Ground Truth Data Annotation, April 2025. 14, 34

[108] Tian Lin, Meng Wang, Aidi Lin, Xiaoting Mai, Huiyu Liang, Yih-Chung Tham, and Haoyu Chen. Efficiency and safety of automated label cleaning on multimodal retinal images. *npj Digital Medicine*, 8(1):10, January 2025. ISSN 2398-6352. doi: 10.1038/s41746-024-01424-x. 63

[109] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. 28, 84

[110] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017. ISSN 1361-8415. doi: 10.1016/j.media.2017.07.005. 34, 103

[111] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, 47(2):519–528, June 2015. ISSN 1554-3528. doi: 10.3758/s13428-014-0483-x. 27, 46, 63

[112] Qian Long, Zhi Li, Ran Gong, Ying Nian Wu, Demetri Terzopoulos, and Xiaofeng Gao. TeamCraft: A Benchmark for Multi-Modal Multi-Agent Systems in Minecraft, December 2024. 38

[113] Xiaolu Lu, David Ratcliffe, Tsu-Ting Kao, Aristarkh Tikhonov, Lester Litchfield, Craig Rodger, and Kaier Wang. Rethinking Quality Assurance for Crowdsourced Multi-ROI Image Segmentation. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1):103–114, November 2023. ISSN 2769-1349. doi: 10.1609/hcomp.v11i1.27552. 31, 32

[114] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):5217, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07619-7. 42, 44, 50, 52, 57

[115] Lena Maier-Hein, Annika Reinke, Michal Kozubek, Anne L. Martel, Tal Arbel, Matthias Eisenmann, Allan Hanbury, Pierre Jannin, Henning Müller, Sinan Onogur, Julio Saez-Rodriguez, Bram van Ginneken, Annette Kopp-Schneider, and Bennett A. Landman. BIAS: Transparent reporting of biomedical image analysis challenges. *Medical Image Analysis*, 66:101796, December 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101796. 18, 50, 55, 58

[116] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M. Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Anna Kisilenko, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Minu D. Tizabi, Matthias Eisenmann, Tim J. Adler, Janek Gröhl, Melanie Schellenberg, Silvia Seidlitz, T. Y. Emmy Lai, Bünyamin Pekdemir, Veith Roethlingshoefer, Fabian Both, Sebastian Bittel, Marc Mengler, Lars Mündermann, Martin Apitz, Annette Kopp-Schneider, Stefanie Speidel, Felix Nickel, Pascal Probst, Hannes G. Kenngott, and Beat P. Müller-Stich. Heidelberg colorectal data set for surgical data science in the sensor operating room. *Scientific Data*, 8(1):101, April 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00882-2. 45, 52, 61, 98

[117] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, Hirenkumar Nakawala, Adrian Park, Carla Pugh, Danail Stoyanov, Swaroop S. Vedula, Kevin Cleary, Gabor Fichtinger, Germain Forestier, Bernard Gibaud, Teodor Grantcharov, Makoto Hashizume, Doreen Heckmann-Nötzel, Hannes G. Kenngott, Ron Kikinis, Lars Mündermann, Nassir Navab, Sinan Onogur, Tobias Roß, Raphael Sznitman, Russell H. Taylor, Minu D. Tizabi, Martin Wagner, Gregory D. Hager, Thomas Neumuth, Nicolas Padoy, Justin Collins, Ines Gockel, Jan Goedeke, Daniel A. Hashimoto, Luc Joyeux, Kyle Lam, Daniel R. Leff, Amin Madani, Hani J. Marcus, Ozanan Meireles, Alexander Seitel, Dogu Teber, Frank Ückert, Beat P. Müller-Stich, Pierre Jannin, and Stefanie Speidel. Surgical data science – from concepts toward clinical translation. *Medical Image Analysis*, 76:102306, February 2022. ISSN 1361-8415. doi: 10.1016/j.media.2021.102306. 6, 28, 29, 40, 94, 103, 107

[118] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, February 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02151-z. 1, 15, 16, 17, 18, 32, 36, 57, 82, 108, 110, 149

[119] Lena Maier-Hein, Annika Reinke, Colliot Olivier, and Bennett Landman. SIG Special Interest Group for

Challenges. https://miccai.org/index.php/special-interest-groups/challenges/sig-members/, September 2025. v, vii, 55, 116

[120] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23803–23828. PMLR, July 2023. 20

[121] Pietro Mascagni, Deepak Alapatt, Aditya Murali, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data*, 12(1):331, February 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-04642-4. 99

[122] Leon Mayer, Tim Rädsch, Dominik Michael, Lucas Luttner, Amine Yamlahi, Evangelia Christodoulou, Patrick Godau, Marcel Knopp, Annika Reinke, Fiona Kolbinger, and Lena Maier-Hein. Challenging Vision-Language Models with Surgical Data: A New Dataset and Broad Benchmarking Study, July 2025. 94, 117, 118

[123] Charles E. McCulloch and Shayle R. Searle. *Generalized, Linear, and Mixed Models*. John Wiley & Sons, 2004. 65

[124] Medical Image Computing and Computer Assisted Intervention. Call for Challenges. http://www.miccai.org/news/2021/10/25/call-for-challenges, October 2021. 52

[125] Milagros Miceli and Julian Posada. The Data-Production Dispositif. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–37, November 2022. ISSN 2573-0142. doi: 10.1145/3555561. 63, 72

[126] Evan Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations, November 2024. 113

[127] George A. Miller. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. 111

[128] Tom Mitchell. Machine Learning. *McCraw Hill*, 1996. 11, 19, 20

[129] Filip Moons and Ellen Vandervieren. Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. A generalisation of Fleiss' kappa. *Behavior Research Methods*, 57(10):287, September 2025. ISSN 1554-3528. doi: 10.3758/s13428-025-02746-8. 33

[130] Kathleen L. Mosier and Linda J. Skitka. Automation Use and Automation Bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 43(3):344–348, September 1999. ISSN 1071-1813. doi: 10.1177/154193129904300346. 113

[131] Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro Von Werra, and Shayne Longpre. OctoPack: Instruction Tuning Code Large Language Models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, November 2023. 38

[132] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. S1: Simple test-time scaling, March 2025. 104

[133] Lalli Myllyaho, Mikko Raatikainen, Tomi Männistö, Tommi Mikkonen, and Jukka K. Nurminen. Systematic Literature Review of Validation Methods for AI Systems. *Journal of Systems and Software*, 181:111050, November 2021. ISSN 01641212. doi: 10.1016/j.jss.2021.111050. 77

[134] Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C. Kot, and Shijian Lu. MMRel: A Relation Understanding Benchmark in the MLLM Era, November 2024. 77

[135] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, April 2021. 12, 23, 29, 31, 55, 57, 149

[136] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, pages 557–566, New York, NY, USA, March 2010. Association for Computing Machinery. ISBN 978-1-60558-815-5. doi: 10.1145/1743384.1743478. 56

[137] Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs, August 2025. 109

[138] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak,

Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. GPT-4o System Card, October 2024. 35

[139] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4): 867–872, 2009. 44

[140] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. 1, 5

[141] Silas Nyboe Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R. Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. A Survey of Crowdsourcing in Medical Image Analysis. *Human Computation*, 7:1–26, December 2020. ISSN 2330-8001. doi: 10.15346/hc.v7i1.1. 4, 11, 30, 105

[142] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, page n71, March 2021. ISSN 1756-1833. doi: 10.1136/bmj.n71. 52

[143] Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games, April 2025. 37

[144] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, November 2021. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100336. 4, 5, 55

[145] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers, November 2021. 55

[146] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C. Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P. Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M. Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M. Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H. Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K. Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, J. P. Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal,

Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah Milbauer, M. Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O, Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J. Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X. Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin,

Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P. V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S. Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben

Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's Last Exam, April 2025. 12, 29, 36, 37, 108

[147] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples, May 2024. 36

[148] Matt Post. A Call for Clarity in Reporting BLEU Scores, September 2018. 36

[149] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. No Filter: Cultural and Socioeconomic Diversity in Contrastive Vision–Language Models. May 2024. 107

[150] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. 24, 31

[151] Tim Rädsch. *The Quest for High Quality Annotations. Quantifying the Impact of Labeling Instructions on Real-World Data*. PhD thesis, Karlsruher Institut für Technologie, Karlsruhe, April 2021. 46

[152] Tim Rädsch, Sven Eckhardt, Florian Leiser, Konstantin D Pandl, Scott Thiebes, and Ali Sunyaev. What Your Radiologist Might be Missing: Using Machine Learning to Identify Mislabeled Instances of X-ray Images. In *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*, 2021. 55

[153] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Heller, Fabian Isensee, Annette Kopp-Schneider, and Lena Maier-Hein. Labeling instructions matter in biomedical image analysis. An annotator-centric perspective. *Medical Imaging meets NeurIPS Workshop*, 2022. 116

[154] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Heller, Fabian Isensee, Annette Kopp-Schneider, and Lena Maier-Hein. Labeling instructions matter in biomedical image analysis. *Medical Imaging meets NeurIPS Workshop*, 2022. 116

[155] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Schreck, A. Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. Labelling instructions matter

in biomedical image analysis. *Nature Machine Intelligence*, 5(3):273–283, March 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00625-5. 1, 2, 11, 12, 13, 14, 26, 28, 31, 41, 116

[156] Tim Rädsch, Leon Mayer, Simon Pavicic, A. Emre Kavur, Marcel Knopp, Barış Öztürk, Klaus Maier-Hein, Paul F. Jaeger, Fabian Isensee, Annika Reinke, and Lena Maier-Hein. Bridging vision language model (VLM) evaluation gaps with a framework for scalable and cost-effective benchmark generation, February 2025. 39, 77, 95, 96, 98, 99, 158

[157] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Heller, Fabian Isensee, Annette Kopp-Schneider, and Lena Maier-Hein. Quality Assured: Rethinking Annotation Strategies in Imaging AI. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15136, pages 52–69. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-73228-7 978-3-031-73229-4. doi: 10.1007/978-3-031-73229-4_4. 27, 28, 59, 116

[158] Rafael Rafailov, Kyle Hatch, Anikait Singh, Laura Smith, Aviral Kumar, Ilya Kostrikov, Philippe Hansen-Estruch, Victor Kolev, Philip Ball, Jiajun Wu, Chelsea Finn, and Sergey Levine. D5RL: Diverse Datasets for Data-Driven Deep Reinforcement Learning, August 2024. 29

[159] Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. BiasDora: Exploring Hidden Biased Associations in Vision-Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10439–10455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.611. 107

[160] Annika Reinke, Matthias Eisenmann, Sinan Onogur, Marko Stankovic, Patrick Scholz, Peter M. Full, Hrvoje Bogunovic, Bennett A. Landman, Oskar Maier, and Bjoern Menze. How to exploit weaknesses in biomedical challenge design and organization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 388–395. Springer, 2018. 52

[161] Annika Reinke, Minu D. Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Emre Kavur, Tim Rädsch, Carole H. Sudre, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Florian Buettner, M. Jorge Cardoso, Veronika Cheplygina, Jianxu Chen, Evangelia Christodoulou, Beth A. Cimini, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, Patrick Godau, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Jens Kleesiek, Florian Kofler, Thijs Kooi, Annette Kopp-Schneider, Michal Kozubek, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Susanne M. Rafelski, Nasir Rajpoot, Mauricio Reyes, Michael A. Riegler, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, Ziv R. Yaniv, Paul F. Jäger, and Lena Maier-Hein. Understanding metric-related pitfalls in image analysis validation. *Nature Methods*, 21(2):182–194, February 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02150-0. 12, 17, 19, 25, 57, 65, 82, 90

[162] Grand View Research. Data Collection And Labeling Market Size & Share Report 2030. https://www.grandviewresearch.com/industry-analysis/data-collection-labeling-market, November 2023. 72

[163] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J. Pappas. Jailbreaking LLM-Controlled Robots, November 2024. 109

[164] Aryelly Rodriguez, Christopher Tuck, Marshall F Dozier, Stephanie C Lewis, Sandra Eldridge, Tracy Jackson, Alastair Murray, and Christopher J Weir. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. *Clinical Trials*, 19 (4):452–463, August 2022. ISSN 1740-7745. doi: 10.1177/17407745221087469. 104

[165] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28. 21

[166] Tobias Roß, Annika Reinke, Peter M. Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Pablo Arbeláez, Gui-Bin Bian, Sebastian Bodenstedt, Jon Lindström Bolmgren, Laura Bravo-Sánchez, Hua-Bin Chen, Cristina González, Dong Guo, Pål Halvorsen, Pheng-Ann Heng, Enes Hosgor, Zeng-Guang Hou, Fabian Isensee, Debesh Jha, Tingting Jiang, Yueming Jin, Kadir Kirtac, Sabrina Kletz, Stefan Leger, Zhixuan Li, Klaus H. Maier-Hein, Zhen-Liang Ni, Michael A. Riegler, Klaus Schoeffmann, Ruohua Shi, Stefanie Speidel, Michael Stenzel, Isabell Twick, Gutai Wang, Jiacheng Wang, Liansheng Wang, Lu Wang, Yujie Zhang, Yan-Jie Zhou, Lei Zhu, Manuel Wiesenfarth, Annette Kopp-Schneider, Beat P. Müller-Stich, and Lena Maier-Hein. Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge. *Medical Image Analysis*, 70:101920, May 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101920. 45, 48, 52, 164

[167] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. 20, 23, 27

[168] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–15, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445518. 1, 2, 4, 56, 57, 74, 108

[169] Cullen Schaffer. Selecting a classification method by cross-validation. *Machine Learning*, 13(1):135–143, October 1993. ISSN 1573-0565. doi: 10.1007/BF00993106. 19

[170] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, January 2015. ISSN 08936080. doi: 10.1016/j.neunet.2014.09.003. 20, 22

[171] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models, October 2022. 25, 103, 105

[172] Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, M. Jehanzeb Mirza, Leshem Chosen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, and Raja Giryes. LiveXiv – A Multi-Modal Live Benchmark Based on Arxiv Papers Content, April 2025. 37

[173] Rohan Shad, John P. Cunningham, Euan A. Ashley, Curtis P. Langlotz, and William Hiesinger. Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. *Nature Machine Intelligence*, 3(11):929–935, November 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00399-8. 55

[174] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020. 31, 56

[175] Sumuk Shashidhar, Clémentine Fourrier, Alina Lozovskia, Thomas Wolf, Gokhan Tur, and Dilek Hakkani-Tür. YourBench: Easy Custom Evaluation Sets for Everyone, April 2025. 109

[176] Ryan Shaw and Patrick Schmitz. Community annotation and remix: A research platform and pilot deployment. In *Proceedings of the 1st ACM International Workshop on Human-centered Multimedia*, HCM '06, pages 89–98, New York, NY, USA, October 2006. Association for Computing Machinery. ISBN 978-1-59593-500-7. doi: 10.1145/1178745.1178761. 13

[177] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017. 37

[178] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, April 2015. 20

[179] LINDA J. Skitka, KATHLEEN L. Mosier, and MARK Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, November 1999. ISSN 1071-5819. doi: 10.1006/ijhc.1999.0252. 30, 113

[180] Elizabeth Stephens. The mechanical Turk: A short history of 'artificial artificial intelligence'. *Cultural Studies*, 37(1):65–87, January 2023. ISSN 0950-2386. doi: 10.1080/09502386.2022.2042580. 27

[181] Norman Makoto Su and David J. Crandall. The Affective Growth of Computer Vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9291–9300, 2021. 110

[182] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. 21

[183] Claude Team. Claude's extended thinking. https://www.anthropic.com/news/visible-extended-thinking, February 2025. 38

[184] Jeyan Thiyagalingam, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. Scientific Machine Learning Benchmarks, October 2021. 16, 17

[185] Luke Tierney. The R statistical computing environment. In *Statistical Challenges in Modern Astronomy V*, pages 435–447. Springer, 2012. 49

[186] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. 74

[187] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36: 46830–46855, 2023. 25

[188] Lukas Tuggener, Jürgen Schmidhuber, and Thilo Stadelmann. Is it enough to optimize CNN architectures on ImageNet? *Frontiers in Computer Science*, 4:1041703, 2022. 110, 111, 154

[189] Lukas Tuggener, Jürgen Schmidhuber, and Thilo Stadelmann. Is it enough to optimize CNN architectures on ImageNet? *Frontiers in Computer Science*, 4:1041703, November 2022. ISSN 2624-9898. doi: 10.3389/fcomp.2022.1041703. 19

[190] Turing. Turing. https://www.turing.com, September 2025. 29

[191] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos, May 2016. 99

[192] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H. S. Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance, October 2024. 108

[193] Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do Large Language Model Benchmarks Test Reliability? https://arxiv.org/abs/2502.03461v1, February 2025. 40

[194] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks, June 2023. 74

[195] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, November 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1724-z. 37

[196] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017. 104, 105

[197] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. Whose AI Dream? In search of the aspiration in data annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–16, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3502121. 26

[198] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks, August 2024. 25, 26

[199] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To See is to Believe: Prompting GPT-4V for Better Visual Instruction Tuning, November 2023. 39, 80, 192

[200] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding, June 2024. 39

[201] Zhecan Wang, Junzhang Liu, Chia-Wei Tang, Hani Alomari, Anushka Sivakumar, Rui Sun, Wenhao Li, Md Atabuzzaman, Hammad Ayyubi, Haoxuan You, Alvi Ishmam, Kai-Wei Chang, Shih-Fu Chang, and Chris Thomas. JourneyBench: A Challenging One-Stop Vision-Language Understanding Benchmark of Generated Images, January 2025. 39, 77, 80, 192

[202] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. 36

[203] David Weslow. Copyright Office Issues Key Guidance on Fair Use in Generative AI Training. https://www.wiley.law/alert-Copyright-Office-Issues-Key-Guidance-on-Fair-Use-in-Generative-AI-Training, 2025. 105

[204] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. 96

[205] Tongyu Wu, James Muldoon, and Bingqing Xia. Global data empires: Analysing artificial intelligence data annotation in China and the USA. *Big Data & Society*, 12(2):20539517251340600, June 2025. ISSN 2053-9517. doi: 10.1177/20539517251340600. 28

[206] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 Technical Report, September 2024. 87, 89, 99

[207] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized Out-of-Distribution Detection: A Survey, January 2024. 111

[208] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2, October 2024. 79, 95

[209] Tianling Yang, Christian Strippel, Alexandra Keiner, Dylan Baker, Alexis Chávez, Krystal Kauffman, Marc Pohl, Caroline Sinders, and Milagros Miceli. Ethics of data work. 2025. 106

[210] Jieping Ye and Jun Liu. Sparse methods for biomedical data. *SIGKDD Explor. Newsl.*, 14(1):4–15, December 2012. ISSN 1931-0145. doi: 10.1145/2408736.2408739. 111

[211] Shuang Yu, Hong-Yu Zhou, Kai Ma, Cheng Bian, Chunyan Chu, Hanruo Liu, and Yefeng Zheng. Difficulty-Aware Glaucoma Classification with Multi-rater Consensus Modeling. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Lecture Notes in Computer Science, pages 741–750, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59710-8. doi: 10.1007/978-3-030-59710-8_72. 33, 40

[212] Kun Yuan, Manasi Kattel, Joel L. Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing Surgical VQA with Scene Graph Knowledge, June 2024. 99

[213] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. HecVL: Hierarchical Video-Language Pretraining for Zero-shot Surgical Phase Recognition, March 2025. 101

[214] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5, 77

[215] Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling Task Transfer Learning, April 2018. 39, 192

[216] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal Neural Script Knowledge Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 23634–23651. Curran Associates, Inc., 2021. 35

[217] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 35

[218] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark, February 2020. 111

[219] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.01100. 24

[220] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, March 2020. ISSN 1941-0484. doi: 10.1109/JSTSP.2020.2987728. 23

[221] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating Agentic Workflow Generation, April 2025. 37

[222] Jieyu Zhang, Le Xue, Linxin Song, Jun Wang, Weikai Huang, Manli Shu, An Yan, Zixian Ma, Juan Carlos Niebles, Silvio Savarese, Caiming Xiong, Zeyuan Chen, Ranjay Krishna, and Ran Xu. ProVision: Programmatically Scaling Vision-centric Instruction Data for Multimodal Language Models, December 2024. 39, 192

[223] Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task Me Anything, January 2025. 39, 192

[224] Rongjian Zhao, Buyue Qian, Xianli Zhang, Yang Li, Rong Wei, Yang Liu, and Yinggang Pan. Rethinking Dice Loss for Medical Image Segmentation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 851–860, November 2020. doi: 10.1109/ICDM50108.2020.00094. 20

[225] X. Zheng, X. Sun, K. Fu, and H. Wang. Automatic Annotation of Satellite Images via Multifeature Joint Sparse Coding With Spatial Relation Constraint. *IEEE Geoscience and Remote Sensing Letters*, 10(4): 652–656, July 2013. ISSN 1558-0571. doi: 10.1109/LGRS.2012.2216499. 31, 34

[226] David Zimmerer, Peter M. Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, Bjørn Sand Jensen, Alison Q. O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, Dmitry V. Dylov, Baolun Yu, Jianyang Zhai, Jingtao Hu, Runxuan Si, Sihang Zhou, Siqi Wang, Xinyang Li, Xuerun Chen, Yang Zhao, Sergio Naval Marimont, Giacomo Tarroni, Victor Saase, Lena Maier-Hein, and Klaus Maier-Hein. MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images. *IEEE Transactions on Medical Imaging*, pages 1–1, 2022. ISSN 1558-254X. doi: 10.1109/TMI.2022.3170077. 44

[227] Eric Zimmermann, Justin Szeto, Jerome Pasquero, and Frederic Ratle. Benchmarking a Benchmark: How Reliable is MS-COCO?, November 2023. 28

[228] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, December 2023. 109

[229] Joseph Friedrich zu Racknitz. *Ueber den Schachspieler des herrn von Kempelen und dessen Nachbildung*. J.G.I. Breitkopf, 1789. 28, 150

Eidesstattliche Versicherung

Statutory Declaration

1. Bei der eingereichten Dissertation zu dem Thema *From Annotation to Validation: Strengthening Machine Learning Pipelines for Real-World Robustness* handelt es sich um meine eigenständig erbrachte Leistung.
   I herewith formally declare that I have written the submitted dissertation From Annotation to Validation: Strengthening Machine Learning Pipelines for Real-World Robustness.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
   I did not use any third party support except for the quoted literature and other sources mentioned in the text. Content from other work, either literally or in content, has been declared as such.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
   The thesis has not been submitted to any examination body in this, or similar, form.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
   I confirm the correctness of the aforementioned declarations.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.
   I am aware of the legal consequences of this declaration. To the best of my knowledge I have told the pure truth and not concealed anything.

Heidelberg, 30.10.2025

Tim Raedsch

List of Acronyms

AI Artificial Intelligence
BERT Bidirectional Encoder Representations from Transformer
CI Credible Interval
CLIP Contrastive Language-Image Pretraining
CNN Convolutional Neural Network
CT Computed Tomography
CVPR Conference on Computer Vision and Pattern Recognition
CVS Critical View of Safety
DKFZ German Cancer Research Center
DL Deep Learning
DSC Dice Similarity Coefficient
ECCV European Conference on Computer Vision
ICCV European Conference on Computer Vision
FDA Food and Drug Administration
FM Foundation Model
FN false negative
FP false positive
GDPR General Data Protection Regulation
GLUE General Language Understanding Evaluation
HeiCo Heidelberg Colorectal
HIT Human Intelligence Task
ICLR International Conference on Learning Representations
IQR Interquartile Range
KI Künstliche Intelligenz
LLM Large Language Model
MCC Matthews Correlation Coefficient
MICCAI Medical Image Computing and Computer Assisted Intervention Society
ML Machine Learning
MMLU Measuring Massive Multitask Language Understanding Benchmark
MMMLU Multilingual Measuring Massive Multitask Language Understanding Benchmark
MTurk Amazon Mechanical Turk
NDA Non-disclosure agreement

NeurIPS Conference on Neural Information Processing Systems
NLP Natural Language Processing
NSD Normalized Surface Distance
ood out-of-distribution
QA Quality Assurance
QAger Qualitätssicherung
RQ Research Question
SOTA state-of-the-art
SVAMP Simple Variations on Arithmetic Math word Problems
USAMO United States of America Mathematical Olympiad
ViT Vision Transformer
VLM Vision-Language Model
VQA Visual Question Answering
VRAM Video Random Access Memory
ZIBMM zero-inflated beta mixed model

List of Figures

List of Tables

A | Appendix

Disclaimer:

The Supplementary contains bloody surgery images that some readers may find disturbing.

A Annotation and Quality Assurance Additional Methods

The appendix figures and tables are part of the respective contribution sections that were part of prior publications. For details refer to the beginning of each contribution section.

A.1 PRISMA Statement

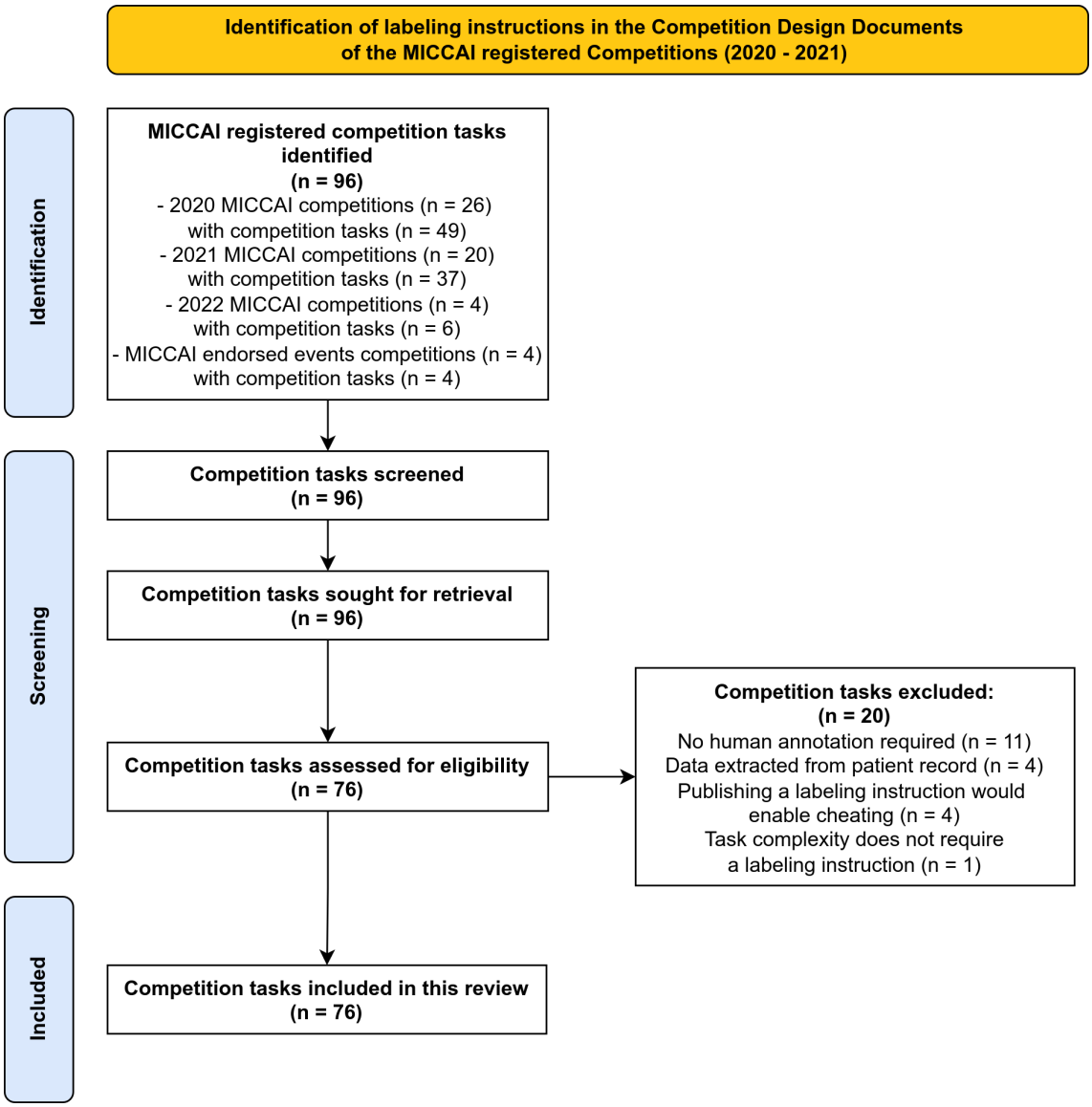


Figure A.1: **Structured identification of labeling instructions in the Challenge Design Documents of the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) registered competitions (2020 - 2021).** After the identification of all relevant competition tasks, each individual competition task was screened and excluded, if the organizers provided a valid reasoning why their labeling instructions cannot be provided.

A.2 Aggregated MICCAI Competition Analysis Results

Table A.1: **Aggregated results of the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) competition analysis.** For each competition, the individual competition task submission documents were evaluated.

Analysis results of MICCAI registered competitions		
Date of evaluation	Nov 10 2021	
Total number of competitions*	53	
Total number of competition tasks	96	
Competition tasks which..	**absolute**	**relative**
do not need to publish LIs or have valid justification**	20	20.83%
Could/should publish	76	79.17%
In scope competition tasks (76) which ...	**absolute**	**relative**
publish LIs	18	23.68%
do not publish / have LIs	58	76.32%
Inter-rater agreement:	90.63%	

** LI = Labeling Instruction

* All information based on the Challenge design docs, mainly question 23 b/a, of the MICCAI registered competitions (Link | accessed: 2021-11-10) between 2020 and 2021.

A.3 Additional Validation Labeling Experiments

Experimental design
High-quality validation/test data is crucial in safety-critical applications. The purpose of this experiment was to investigate the performance variability resulting from different labeling instructions and annotation providers. To this end, we leveraged the existing output of six algorithms that participated in the instance segmentation task of the Robust Medical Instrument Segmentation (ROBUST-MIS) challenge [166] and that were trained on expert-generated annotations (n = 5,983). For each annotated image in our labeled dataset (n = 14,040) introduced in the Method section, we proceeded as follows: We regarded each annotated image obtained with any of the three labeling instruction types and any of the five annotation providers as reference annotation for a tuple (annotation provider, labeling instruction). This resulted in 3 (types of instructions) x 5 (annotation providers) x 4 (annotations per image and provider) = 60 different annotation masks in total for each unique image, as detailed in the Methods section. These new reference annotations and the original annotations, generated by experts with what we henceforth call expert labeling instruction type, were all compared to the fixed model outputs of the six algorithms. We then assigned a "valid" label to all images in which no false positive (FP) or false negative (FN) instruments occurred. For all valid annotations, the mean DSC, averaged over all instances, was determined. This yielded a total of 85,644 valid/invalid scores / 54,475 DSC image scores (for the valid images) with corresponding meta information encoding the image id, the algorithm, the labeling instruction type, the annotation provider and the annotation provider type. Based on the generated data, mixed model analysis enabled us to quantify the performance variability resulting from different labeling instruction types and annotation providers. Specifically, we used a two-part beta mixed model which included a) a logistic mixed model analyzing the probability of valid annotations and b) a beta mixed model analyzing the non-zero DSC values of an image when only valid annotations occurred. The image id and the fixed algorithm outputs were defined as random effects. We defined the labeling instruction type and the annotation provider type as fixed effects, in which we are interested, with the minimal text instruction type and crowdworkers as reference categories, respectively. This enabled us to calculate the effects of these two variables on both the probability of valid annotations and on the DSC scores.

The model was implemented with R (version 4.0.3, packages: lme4 version 1.1.27.1, glmmTMB version 1.1.2).

Results
Figure A.2 shows the effects of the labeling instruction type on (a) the probability of valid annotations and (b) the effect on the DSC scores. In comparison to the valid annotation odds of the minimal text labeling instruction type, the valid annotation odds were substantially higher for the expert labeling instruction type and the picture labeling instruction type, and lower for extended text labeling instruction type (cf. Figure A.2a). We observe a comparable change in the ratio between the expected DSC score and the difference to a perfect DSC score for the expert labeling instruction type with an increase of 6% (1.06, Credible Interval (CI): 1.01, 1.10) and a reduction of 3% (0.97, CI: 0.96, 0.99) for the extended text labeling instruction type, in comparison to the minimal text labeling instruction type, once an object was identified. Interestingly, we observed a minor reduction, which was not statistically significant, in the ratio between the expected DSC score and the difference to a perfect DSC score with the picture labeling instruction type, as depicted in Figure A.2b. Under the same conditions, the odds of valid annotations for professional annotators were 2.17 times (CI: 2.08, 2.27) that of crowdworkers, and, for experts, 2.22

times (CI: 1.93, 2.56) that of crowdworkers. Similarly, in comparison to crowdworkers, the ratio between the expected DSC score and the difference to a perfect DSC score for the same predicted images by the different algorithms were increased by 39% (1.39, CI: 1.37, 1.41) for professional annotators and by 41% (1.41, CI: 1.35, 1.48) for the experts. **Thus, test set images annotated by professional annotation companies without additional quality assurance are very similar in annotation quality to expert-annotated test set images with quality assurance for the task at hand.**

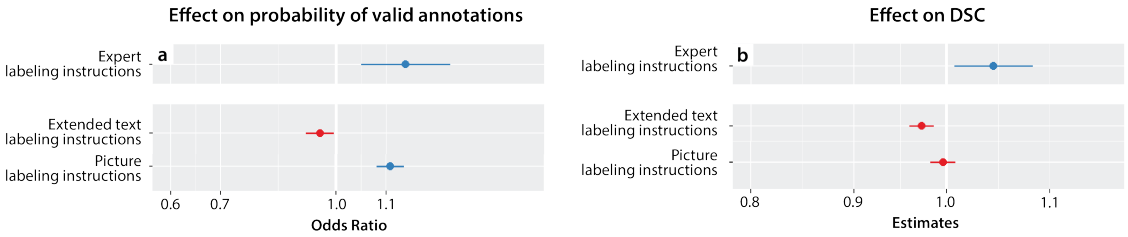


Figure A.2: **Effect of the labeling instruction and the annotator provider type on validation/testing:** The experiments indicate that poor choices of annotators and labeling instructions systematically lead to misreporting the performance of machine learning (ML) models on the test set. (**a**) For the exact same predictions of ML models, extended text descriptions reduce the odds for valid annotations in comparison to minimal text descriptions, while the expert labeling instruction type and including pictures provide a clear benefit. (**b**) Only the expert labeling instructions, which were exclusively used in the 2019 Robust Medical Instrument Segmentation challenge, show an improvement on the Dice Similarity Coefficient (DSC) score. The change (increase: blue, decrease: red) of the valid annotations probability (a) and the perfect DSC score odds (b) is displayed with the minimal text labeling instruction as reference. The two-part beta model was fit on 85,644 DSC image scores from the existing output of six different ML models, our labeled dataset and the expert generated images.

A.4 Survey: Current Status of Labeling Instructions

Labeling Instructions Survey

Section 1: Introduction

This survey aims to collect information about people who perform annotations. Labeling instructions provide the information on "what and how to annotate".

It will take around 5 minutes to finish the survey.
The collected information will help us improve labeling instructions in the future.
The collected information is anonymous and will not be shared with your employer.

This survey consists of two parts:
- General questions
- Annotation-related questions

contact information: xxxx

I hereby confirm that my responses will be processed in an anonymous fashion in a study on labeling instructions.

- Yes

Section 2: General questions

What device are you using for the labeling?

Question type: Multiple choice

- Desktop PC
- Laptop

How old are you?

Question type: Drop-down

- 15-20 years
- 21-30 years
- 31-40 years
- 41-50 years
- 50-60 years
- 61-70 years
- Other

Please rate your English skills.

Question type: Drop-down

- Beginner
- Elementary
- Intermediate

- Advanced
- Proficient
- Native speaker

What is your native language?

Question type: Open text

➢ YOUR ANSWER

How long have you been using a computer for?

Question type: Drop-down

- 0-5 years
- 6-10 years
- 11-15 years
- 16-20 years
- More than 20 years

What is your highest earned degree?

Question type: Drop-down

- Basic education
- Higher education (similar to high school)
- Bachelor - University or similar
- Master - University or similar
- Doctor - University or similar
- None of the above

Section 3: Annotation-related questions

How often do you label per week?

Question type: Drop-down

- 1 day per week
- 2 days per week
- 3 days per week
- 4 days per week
- 5 days per week
- More than 5 days per week

Which of the mentioned options cause problems in the daily annotation work of your colleagues? You can choose multiple answers.

Question type: Checkboxes

- Missing training
- Unclear labeling instructions / Unclear labeling guide
- Tooling issues / Tooling restrictions (e.g. not able to save annotations)
- Poor data (e.g. image quality or sensor fusion)
- Concentration issues (due to monotonous work)
- Other…

Is labeling your main source of income?

Question type: Multiple choice

- Yes
- No

How many hours do you spend per week on labeling images? Please note that all activities regarding labeling (e.g. training) are counted here.

Question type: Drop-down

- 1-2 hours
- 3-5 hours
- 6-10 hours
- 11-15 hours
- 16-20 hours
- 21-30 hours
- 31-40 hours
- more than 40 hours

How long have you been labeling for?

Question type: Drop-down

- Less than 1 year
- Between 1 and 1.5 years
- Between 1.6 and 2 years
- Between 2.1 and 2.5 years
- Between 2.6 and 3 years
- Between 3.1 and 4 years
- Between 4 and 5 years
- Between 6 and 7 years
- Between 8 and 9 years
- More than 10 years

Please rate the following statements. If you do not understand the statement, you can select "Do not understand statement" for certain statements.

I prefer short labeling instructions rather than long labeling instructions.

Question type: Multiple choice (Likert Scale)

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

Labeling instructions are not important for annotation projects.

Question type: Multiple choice (Likert Scale)

- Strongly disagree
- Disagree
- Neither agree nor disagree

- Agree
- Strongly agree

After reading more detailed labeling instructions my colleagues are able to generate annotations faster than with less detailed labeling instructions.

Question type: Multiple choice (Likert Scale)

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree
- Do not understand statement

After reading more detailed labeling instructions my colleagues are able to generate annotations with higher quality than with less detailed labeling instructions.

Question type: Multiple choice (Likert Scale)

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree
- Do not understand statement

The creator of the labeling instructions should invest more time and resources in the generation of the labeling instructions.

Question type: Multiple choice (Likert Scale)

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree
- Do not understand statement

Please rate the following statements. The scale options are defined as:
Never: 0%
Rarely: Between 1% and 25%
Sometimes: Between 26% and 50%
Often: Between 51% and 75%
Always: Between 76% and 100%

Unclear labeling instructions lead to questions of my colleagues about the labeling instructions.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes

- Often
- Always

In the past, my colleagues made annotation mistakes due to unclear labeling instructions.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes
- Often
- Always

Extensive text descriptions and written examples in the labeling instructions improve the understanding of the labeling instructions for my colleagues.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes
- Often
- Always

In the past, my colleagues made annotation mistakes due to incomplete labeling instructions.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes
- Often
- Always

Pictures in the labeling instructions improve the understanding of the labeling instructions for my colleagues.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes
- Often
- Always

In the past, I made annotation mistakes due to unclear labeling instructions.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes
- Often
- Always

Please rate the following statements. If you do not understand the statement, you can select "Do not understand statement" for certain statements. The scale options are defined as:
Never: 0%
Rarely: Between 1% and 25%
Sometimes: Between 26% and 50%
Often: Between 51% and 75%
Always: Between 76% and 100%

The labeling instructions my colleagues receive for annotation projects are usually of good quality and enable them to generate high quality annotations.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes
- Often
- Always
- Do not understand statement

Unclear labeling instructions delay the completion of annotation projects.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes
- Often
- Always
- Do not understand statement

Unclear labeling instructions cause rework of the annotations for my colleagues.

Question type: Multiple choice (Likert Scale)

- Never
- Rarely
- Sometimes
- Often
- Always
- Do not understand statement

Is there anything you would like to add or mention?

Question type: Open text

➢ YOUR ANSWER

Thank you for participating in the survey and helping to improve labeling instructions.

A.5 Minimal Text Labeling Instruction

Overview

1

Terminology

Definition “Matter”:

- Anything that has mass, takes up space and can be clearly identified.
- Examples: tissue, surgical tools, blood

2

Goal - Detect and segment all relevant medical instruments

Only segment medical instruments

- Each medical instrument is their own instance.
- Each instance has their own colour.
- The interior of a segmentation represents a medical instrument.
- Everything outside the medical instrument segmentations is regarded as other matter.

3

Occlusion

Each pixel may correspond to exactly one instance.
The solid/liquid matter that occurs first along the line of sight of the endoscope determines the label.

4

Medical instruments

Definition “Medical instrument”:

- An elongated rigid object placed inside the patient and manipulated from outside the patient.
- Examples: grasper, scalpel, (transparent) trocar, clip applicator, hooks, stapling device, suction

5

Medical instruments - Holes

Several medical instruments feature holes.
They are regarded as part of the instrument.

6

Medical instruments - Transparency

Medical instruments may be transparent.
The occlusion rule holds in this case as well.

A.6 Extended Text Labeling Instruction

Overview

1

Terminology

Definition “Matter”:

- Anything that has mass, takes up space and can be clearly identified.
- Examples: tissue, surgical tools, blood
- Counterexamples: reflections, digital overlays, movement artifacts, smoke

2

Goal - Detect and segment all relevant medical instruments

Only segment medical instruments

- Each medical instrument is their own instance.
- If two parts of the same instrument are visible, they belong to the same instrument instance.
- Each instance has their own colour.
- The interior of a segmentation represents a medical instrument.
- Everything outside the medical instrument segmentations is regarded as other matter.
- Be as precise as possible. Every pixel counts in surgery.

3

Occlusion

Each pixel may correspond to exactly one instance.
The solid/liquid matter that occurs first along the line of sight of the endoscope determines the label.

This may result in multiple segmentations for a single instrument that is occluded by another instrument, blood or tissue, for example.

4

Medical instruments

Definition “Medical instrument”:

- An elongated rigid object placed inside the patient and manipulated from outside the patient.
- Examples: grasper, scalpel, (transparent) trocar, clip applicator, hooks, stapling device, suction
- Counterexamples: non-rigid tubes, bandage, compress, needle (not directly manipulated from outside but manipulated with an instrument), coagulation sponges, metal clips

5

Medical instruments - Holes

Several medical instruments feature holes.
They are regarded as part of the instrument.

A hole is made up of pixels that do not show

parts of the instrument but are either

type a) completely surrounded by pixels of the same instrument or

type b) completely surrounded by pixels of one instrument and the margin of the instrument would close the hole outside the image.

6

Medical instruments - Holes exception

The sole exception are trocars when the camera is placed inside them.

Trocars are transparent medical instruments

- They are placed through the abdomen.
- They function as a portal for the other medical instruments.
- When the camera is inside a trocar, a trocar can take up a large part of the image.

Trocars are exempt from the hole rule

- Reason: Otherwise, the holes would make up a large part of the image.

7

Medical instruments - Transparency

Medical instruments may be transparent.
The occlusion rule holds in this case as well.

8

Text overlay

Text overlay is text that is visible in the image.
The background of the text is the regular image.

Text overlay shall be ignored.

9

Image overlay

Image overlay is often uniform coloured shapes
that overlay with the image.
Text with an added uniform color background
is considered as image overlay.

Image overlay is treated like “other matter” and should not be part of the segmentation.

10

A.7 Extended Text Including Pictures Labeling Instruction

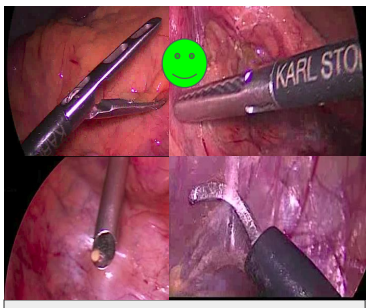
Overview

1

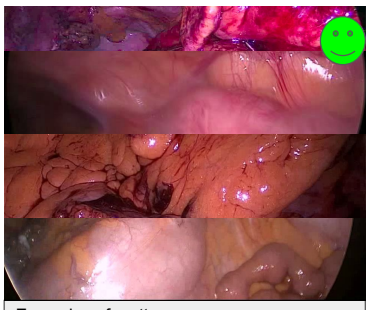
Terminology

Definition "Matter":

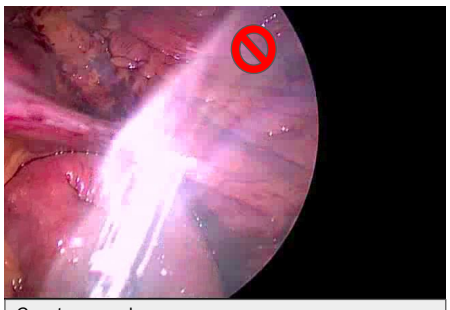
- Anything that has mass, takes up space and can be clearly identified.
- Examples: tissue, surgical tools, blood
- Counterexamples: reflections, digital overlays, movement artifacts, smoke



Examples of matter:
Surgical tools.



Examples of matter:
Tissue.



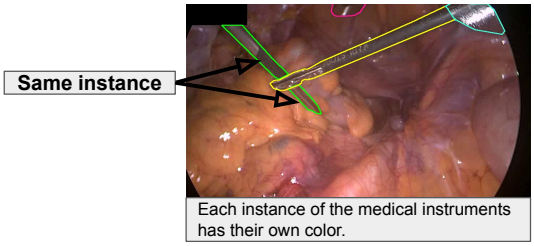
Counterexample:
Smoke.

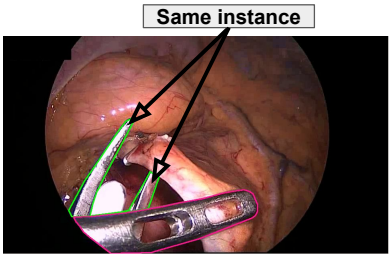
2

Goal - Detect and segment all relevant medical instruments

Only segment medical instruments

- Each medical instrument is their own instance.
- If two parts of the same instrument are visible, they belong to the same instrument instance.
- Each instance has their own colour.



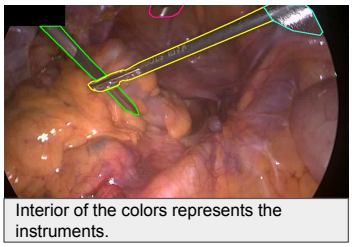


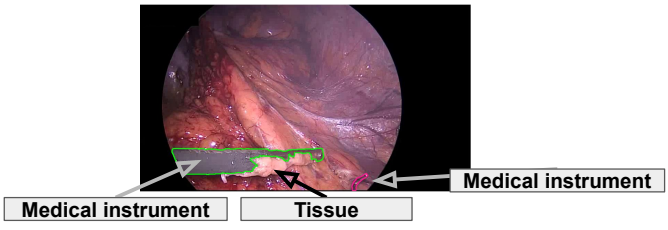
3

Goal - Detect and segment all relevant medical instruments

Only segment medical instruments

- The interior of a segmentation represents a medical instrument.
- Everything outside the medical instrument segmentations is regarded as other matter.
- Be as precise as possible. Every pixel counts in surgery.



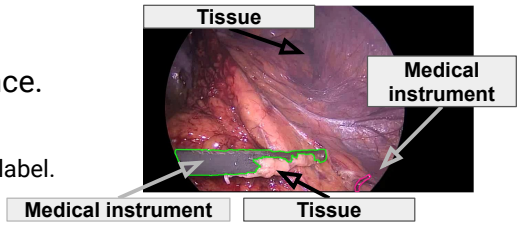


4

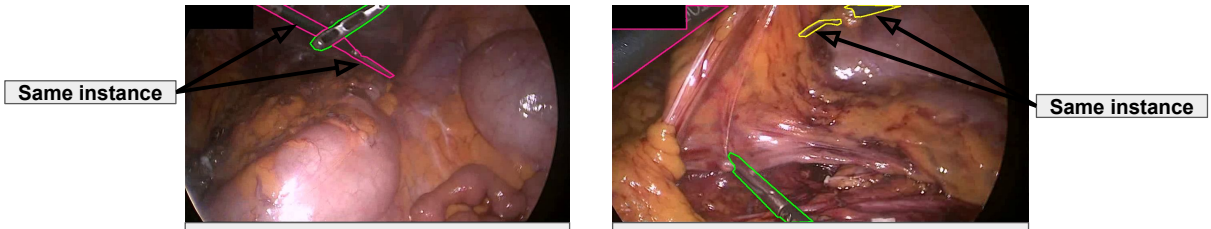
Occlusion

Each pixel may correspond to exactly one instance.

- The solid/liquid matter that occurs first
  along the line of sight of the endoscope determines the label.



This may result in multiple segmentations for a single instrument that is occluded by another instrument, blood or tissue, for example.



Pink is a medical instrument and occluded by green. -> Pink has two segmentations.

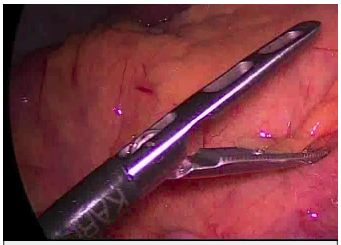
Pink is partly occluded by an image overlay. Yellow is occluded by tissue. -> Two segmentations.

5

Medical instruments

Definition “Medical instrument”:

- An elongated rigid object placed into the patient and manipulated from outside the patient.
- Examples: grasper, scalpel, (transparent) trocar, clip applicator, hooks, stapling device, suction



Example: Grasper.

Example: Grasper.

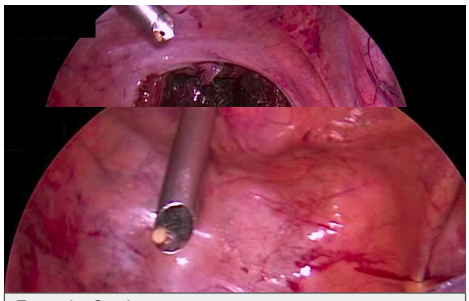
Example: Grasper.

6

Medical instruments

Definition “Medical instrument”:

- An elongated rigid object placed into the patient and manipulated from outside the patient.
- Examples: grasper, scalpel, (transparent) trocar, clip applicator, hooks, stapling device, suction



Example: Suction.

Example: Scalpel.

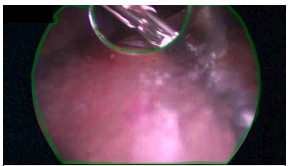
Example: Spreader.

7

Medical instruments

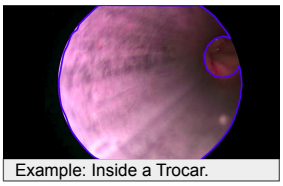
Definition “Medical instrument”:

- An elongated rigid object placed into the patient and manipulated from outside the patient.
- Examples: grasper, scalpel, (transparent) trocar, clip applicator, hooks, stapling device, suction

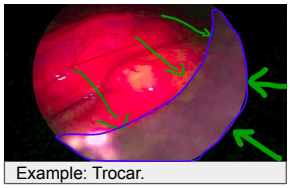


Example: Inside trocar (green). Another instrument is visible through the hole.

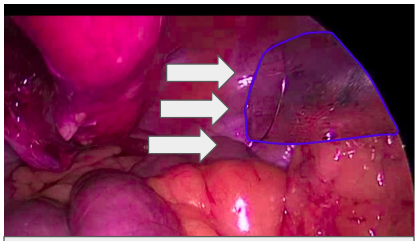
Example: Different trocars.



Example: Inside a Trocar.



Example: Trocar.



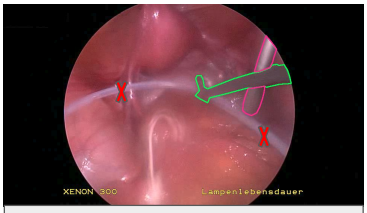
Example: transparent trocar (hard to detect).

8

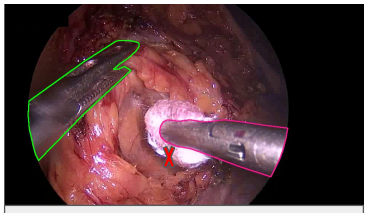
Medical instruments

Definition "Medical instrument":

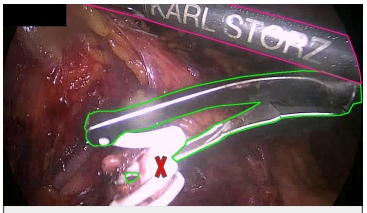
- Counterexamples: non-rigid tubes, bandage, compress, needle (not directly manipulated from outside but manipulated with an instrument), coagulation sponges, metal clips



Green and pink are medical instruments.
The red cross marks a drain-plastic
-> no medical instrument.



Green and pink are medical instruments.
The red cross marks a bandage
-> no medical instrument.



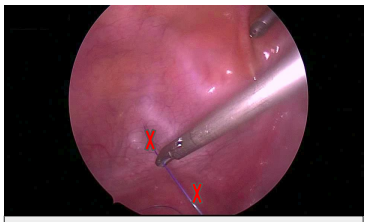
Green and pink are medical instruments.
The red cross marks a clamp
-> no medical instrument.

9

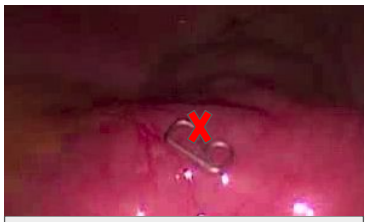
Medical instruments

Definition "Medical instrument":

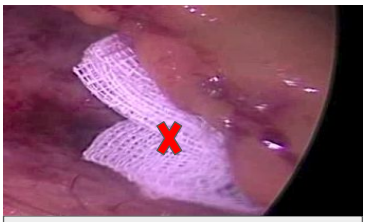
- Counterexamples: non-rigid tubes, bandage, compress, needle (not directly manipulated from outside but manipulated with an instrument), coagulation sponges, metal clips



2 instruments are visible.
The red cross marks a needle
-> no medical instrument.



0 instruments are visible.
The red cross marks a metal clip
-> no medical instrument.



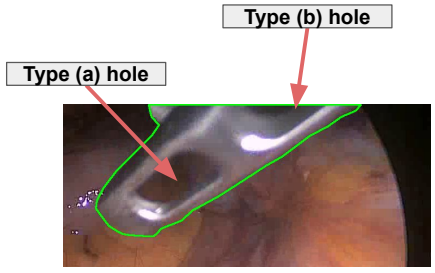
0 instruments are visible.
The red cross marks a bandage
-> no medical instrument.

10

Medical instruments - Holes

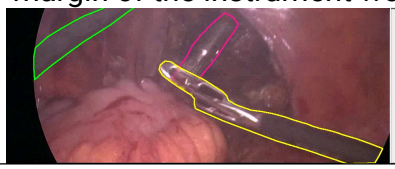
Several medical instruments feature holes. They are regarded as part of the instrument.

A hole is made up of pixels that do not show parts of the instrument but are either



type (a) completely surrounded by pixels of the same instrument or

type (b) are completely surrounded by pixels of one instrument and the margin of the instrument would close the hole outside the image.



Type (a) hole for yellow -> the hole belongs to yellow. Even instruments that are visible in the hole (pink) are ignored.

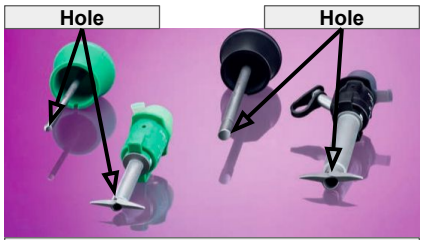
11

Medical instruments - Holes exception

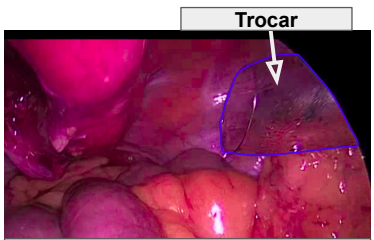
The sole exception are trocars when the camera is placed inside of them.

Trocars are transparent medical instruments.

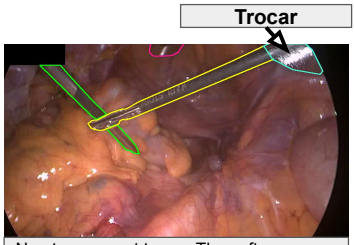
- They are placed through the abdomen.
- To function as a portal for the other medical instruments.



Different types of trocars. They function as a tunnel for other instruments.



Trocars are often transparent and hard to spot.



Non transparent trocar. They often appear at the edge of the image.

12

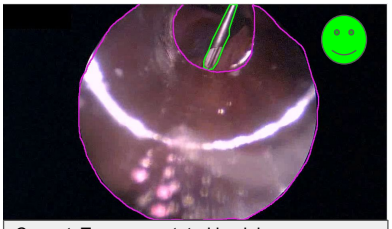
Medical instruments - Holes exception

The sole exception are trocars when the camera is placed inside of them.

When the camera is inside a trocar, a trocar can take up a large part of the image

Trocars are exempt from the hole rule.

- Otherwise, the holes would contain a large part of the image.



Correct: Trocar annotated in pink.

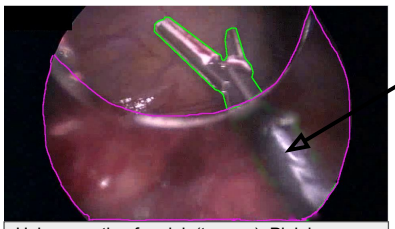
Wrong: Trocar annotated in red.

13

Medical instruments - Transparency

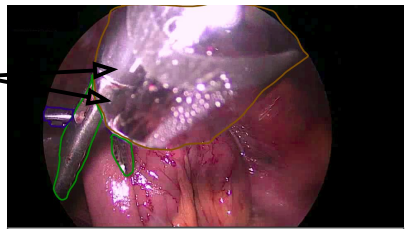
Medical instruments may be transparent.

The occlusion rule holds in this case as well.



Hole exception for pink (troncar). Pink is transparent. -> The hidden part of green belongs to the pink annotation (occlusion rule).

Pixels only belong to one instrument. Here they belong to the transparent instrument, because the transparent instrument is the first object in line of sight (occlusion rule).



Green and blue are instruments. Brown is a transparent trocar. -> The trocar is the first part in the line of sight. Hidden part of the green instrument belongs to brown (occlusion rule).

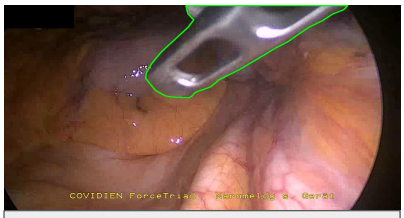
14

Text overlay

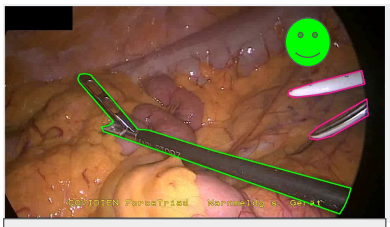
Text overlay is text that is visible in the image.
The background of the text is the regular image.

Text overlay shall be ignored.

Examples:



Text overlay (in yellow color).



Correct: The text overlay is ignored.

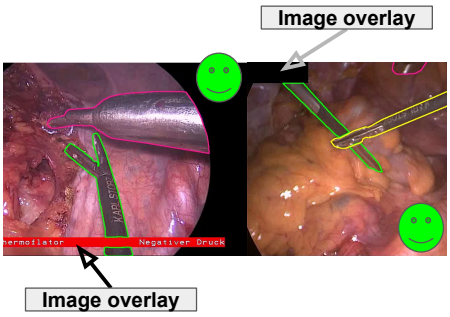
Wrong: The text overlay is treated with a cutout.

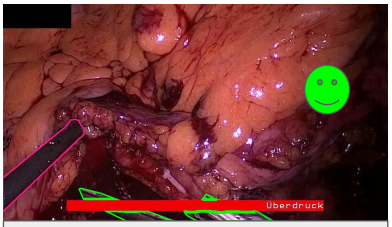
15

Image overlay

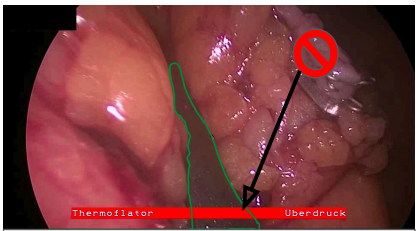
Image overlay is often uniform coloured shapes that overlay on the image.
Text with a added uniform color background is considered as image overlay.

Image overlays are treated like “other matter”.





Correct: The image overlay is not part of the instrument.



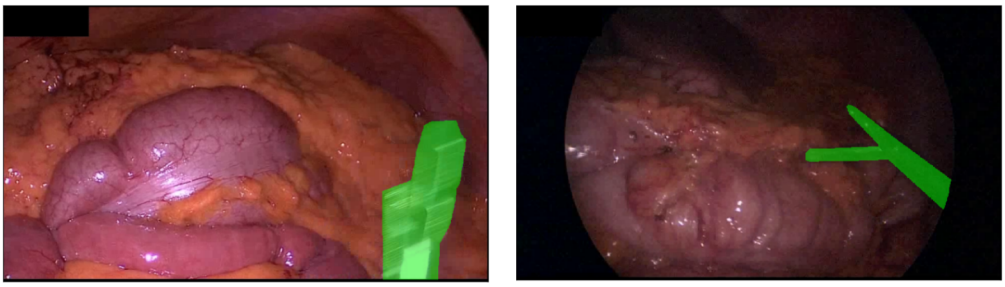
Wrong: The image overlay here is part of the instrument.

16

A.8 Real-world Image Characteristics Examples

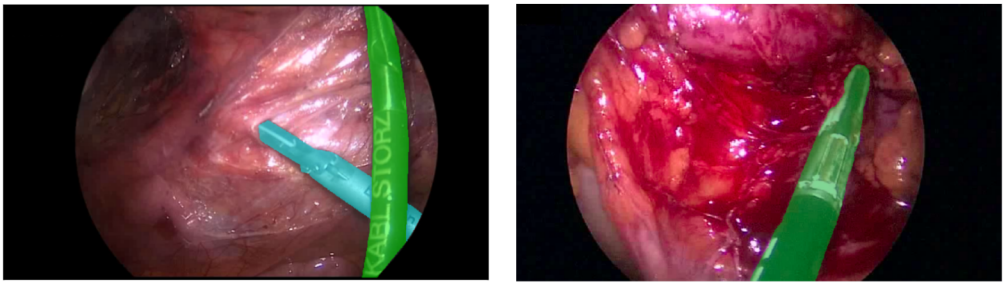
The 57,636 metadata annotations represent real-world image characteristics and were created by three engineers with the help of some crowdworkers. Examples are displayed in Supplementary Fig. A.3. The characteristics can be generalizable (Supplementary Fig. A.3a) or domain-specific (Supplementary Fig. A.3b), and occur in the background, on the objects or as an overlay. Generalizable characteristics include: motion artifacts in the background, other objects present in the background, object occluded by background, intersection of objects, motion artifacts on the object, object(s) overexposure, object(s) underexposure. Domain-specific characteristics include: background covered by blood, background covered by reflections, background covered by smoke, image overlaid with text, caption at the bottom of the image, black box present in the image, trocar present in the image, object covered by blood, object(s) covered by reflections, object(s) covered by smoke.

a Generalizable metadata characteristics



i Motion artifacts on the object

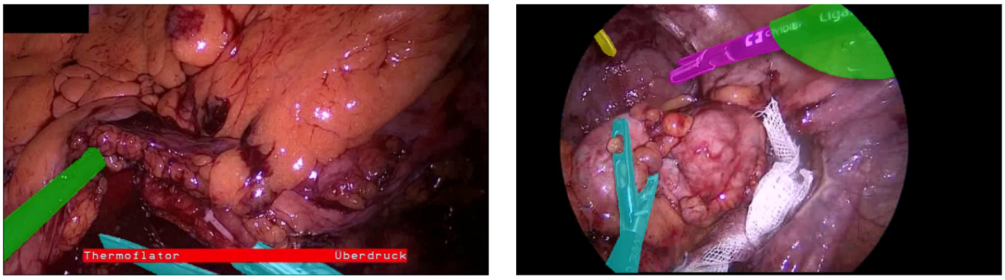
ii Object(s) underexposure



iii Intersection of objects

iv Object occluded by background

b Domain-specific metadata characteristics



i Caption at the bottom of the image

ii Trocar present in the image

Figure A.3: **Real-world image characteristics examples.** During recording, a broad range of (a) generalizable and (b) domain-specific image characteristics occur and can influence humans and models alike.

A.9 Spam Annotation Examples

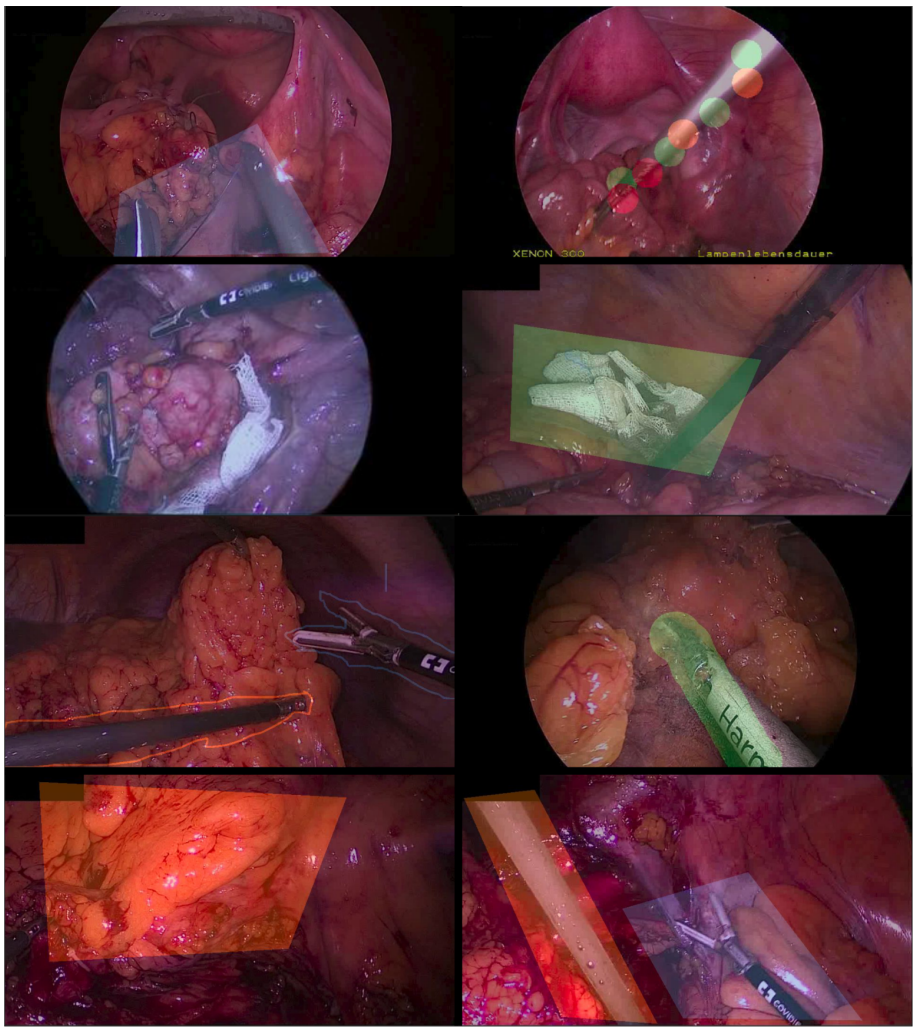


Figure A.4: **Only MTurk crowdworkers generated spam annotations.** The annotation companies generated no spam annotations, whereas around 20% of the collected MTurk annotations were spam. The spam annotations vary greatly in their style and seem to display different strategies to game the MTurk annotation tasks.

B DomainBench Additional Methods

B.1 Task Augmentation Methods Comparison

Metric	CLEVR	Task Me Anything	Taskonomy	Wang 2023	JourneyBench	ProVision	Ours
	[82]	[223]	[215]	[199]	[201]	[222]	**Ours**
Real images/objects	❌	☑️ (partly, needs scene graph)	✅	✅	❌	✅	✅
Diversity core perception tasks	✅ (subjective)	✅	✅	✅	✅	✅	✅
Focus on resource efficiency	Synthetic data	Strong synthetic data focus and flexible	Not relevant, no new tasks/data can be added	❌	❌ (2,200 hours of human annotation)	✅	✅
Enables others to use their own data and benchmark / Easily extendable	✅	✅	❌	✅ (in theory possible, but no code)	✅	✅	✅
Not reliant on generative models	✅	✅	✅	❌	❌	❌	✅
Object-centric	✅	✅	❌	✅	❌ (hard to say)	✅	✅
Validated across multiple visual content domains	❌	❌	❌	✅	✅	✅	✅
Human Ambiguity Scores	❌	❌	❌	❌	❌	❌	✅
Easily scalable	✅	✅	❌	✅	✅	✅	✅
Task creation code public	✅	✅	❌	❌	✅	✅	✅
Evaluated on SOTA VLMs	❌	☑️(limited # of proprietary models)	❌	☑️ (but SoTA* 2023)	☑️(limited # of proprietary models)	❌	✅

Table A.2: **Our framework uniquely combines comprehensive evaluation capabilities with methodological advantages.** Systematic comparison of task augmentation approaches across key metrics, highlighting distinct features in VLM evaluation, programmatic task generation, and efficiency measures relative to existing frameworks.

Our work is positioned within the broader context of research on large VLMs, programmatic task generation and task augmentation. A comparison to other relevant work [82, 223, 215, 199, 201] is provided in Table A.2.

We deliberately excluded large datasets primarily designed for task generation, such as Visual Genome [93] and the Open Images Dataset [96], as these datasets are not suitable for directly using their own images for evaluation. Instead, our focus is on empowering domain-specific users to efficiently test a wide range of capabilities with a small number of real-world images containing relevant objects, while operating under limited resources. For instance, the human annotations required to generate the tasks in our framework can be achieved at a minimal cost of approximately $27 USD (average across all seven datasets). Our approach is complementary to existing research, offering a lightweight and accessible alternative for specific use cases.

Thus enabling researchers and practitioners to iterate more quickly and effectively in the right direction.

B.2 Vision-Language Model Tasks Overview

Here we present the VLM tasks overview and its corresponding meta categories in Figure A.5. Further information on each task is provided in Table 4.7 on the next page.

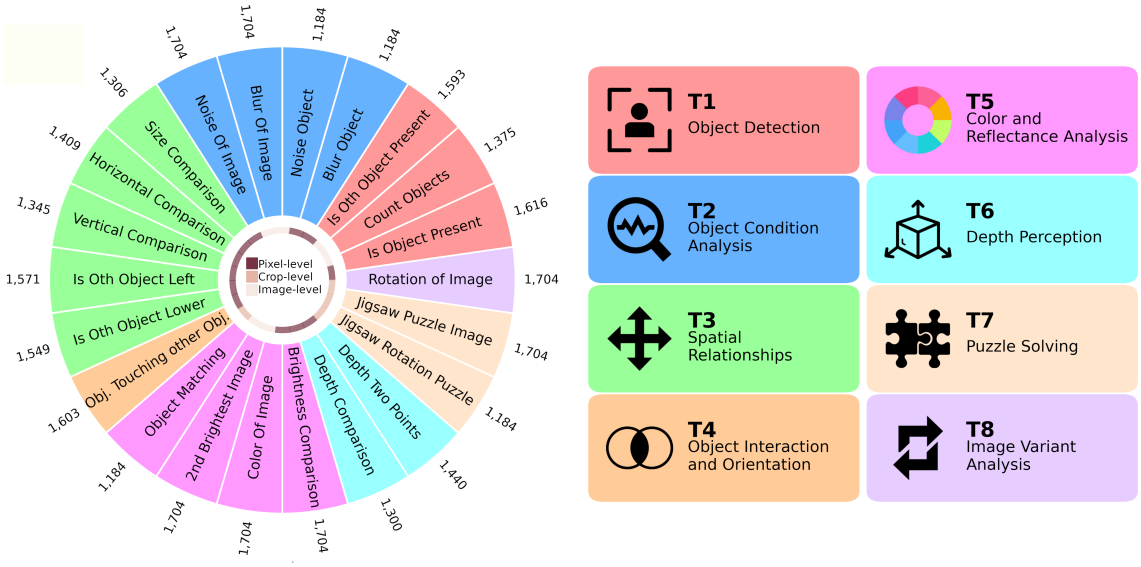


Figure A.5: **Our framework yields a diverse set of tasks.** Based on a single image with instance segmentations, our framework enables the generation of 22 tasks from eight different vision-language categories, ranging from pixel-level to image-level perception.

B.3 Computational and Monetary Cost

Domain	Icon	Metadata Enrichment		Perception Task Generation (in seconds)
		Pre-defined Heuristics (in seconds)	ML Model (in seconds)	
Wildlife	🦁	15	50	124
Persons	🧍	354	496	160
Vehicles	🚘	73	211	129
Animals	🐕	41	138	129
Kitchen	🍳	48	156	161
Food	🥦	295	117	144
Kitti	🐱	23	108	146

Table A.3: **Our framework is resource-efficient in both computation and cost.** Displayed are the timings (in seconds) for metadata enrichment and question generation across domains, demonstrating the scalability and efficiency of our approach.The turnover time for task annotations was two days. All computations were performed on an RTX 3090 GPU and Ryzen 9 5900X CPU.

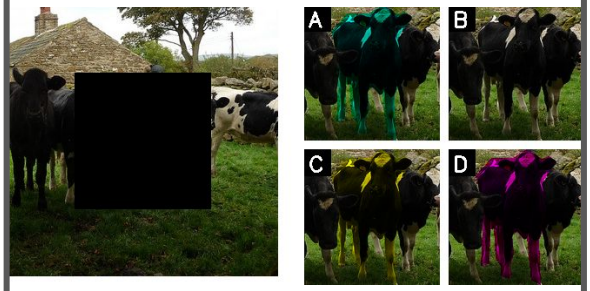
C Domain-specific Datasets

In the following pages, we present examples for all generated datasets. These cover a diverse range from animals up to kitchen scenes.

C.1 Animals Dataset Examples

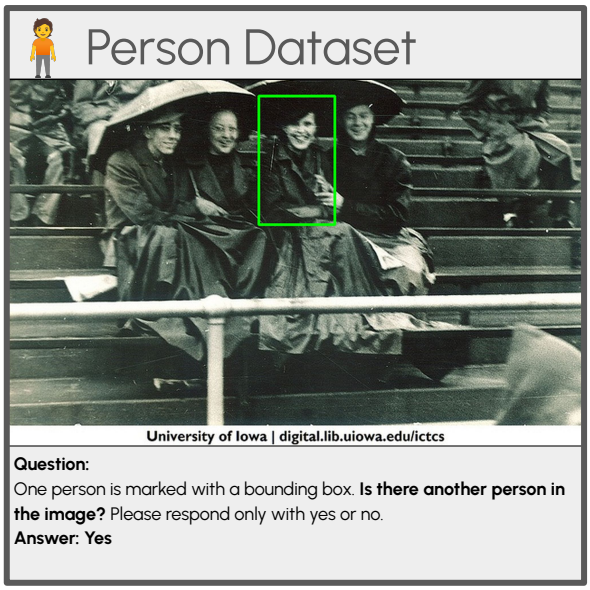
Question:
Which point is brighter? Please only respond with the letter.
Answer: B

Question:
Which tile fits best in the image? Please respond only with the letter..
Answer: A

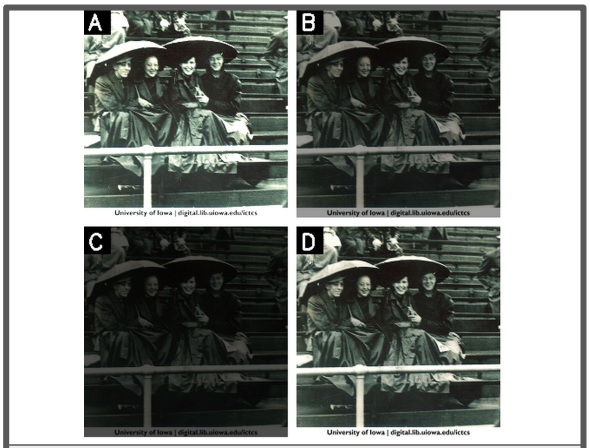


Question:
Which tile shows the correct color pattern? Please respond only with the letter.
Answer: B

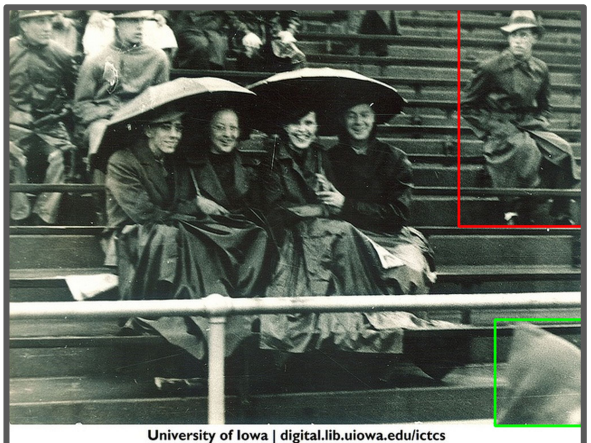
C.2 Person Dataset Examples



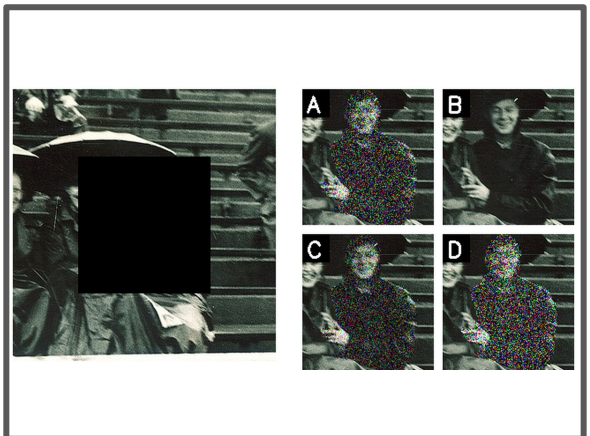
Question:
One person is marked with a bounding box. **Is there another person in the image?** Please respond only with yes or no.
Answer: Yes



Question:
Which image variant is the second brightest version of the image? Please respond only with the letter.
Answer: D



Question:
Each person is marked with a coloured bounding box. **Which person is closer to the camera?** Only respond with the color of the bounding box.
Answer: Green

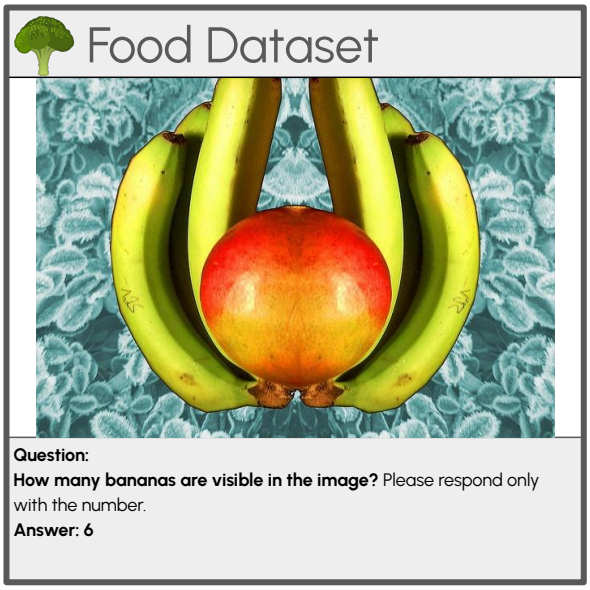


Question:
Which tile fits best in the image? Please respond only with the letter.
Answer: B

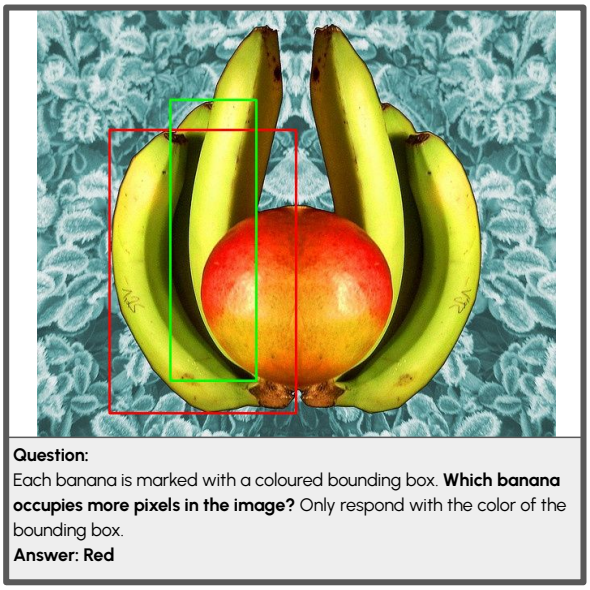
University of Iowa | digital.lib.uiowa.edu/ictcs

Question:
Which point is brighter? Please respond only with the letter.
Answer: A

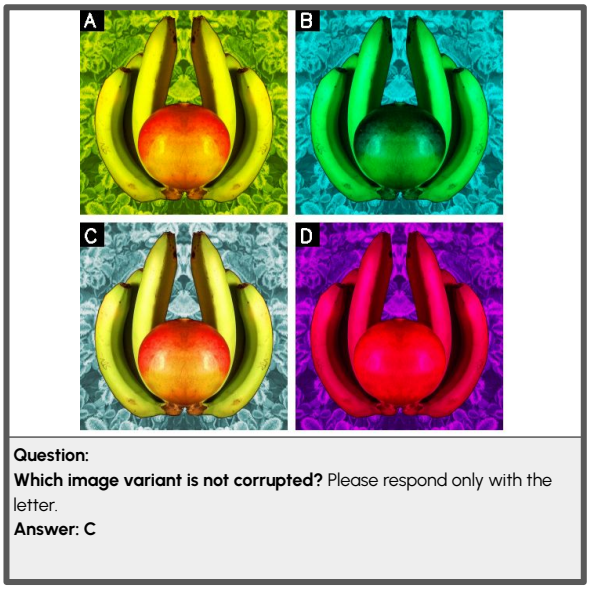
C.3 Food Dataset Examples



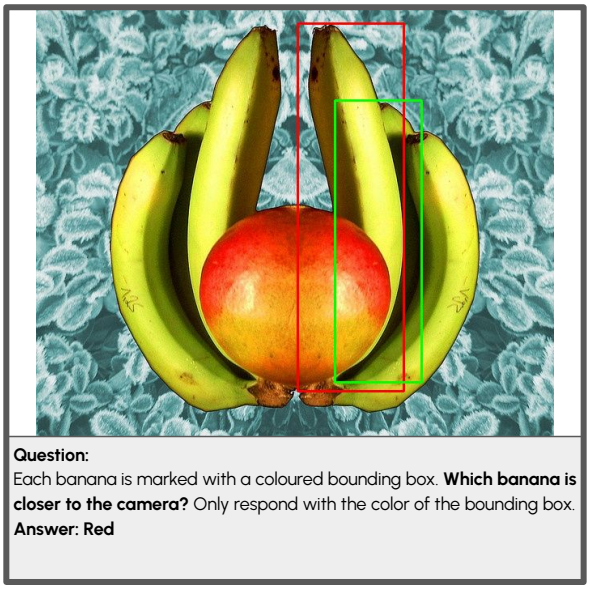
Question:
How many bananas are visible in the image? Please respond only with the number.
Answer: 6



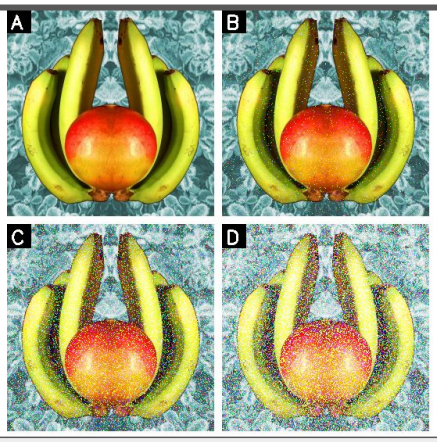
Question:
Each banana is marked with a coloured bounding box. **Which banana occupies more pixels in the image?** Only respond with the color of the bounding box.
Answer: Red



Question:
Which image variant is not corrupted? Please respond only with the letter.
Answer: C



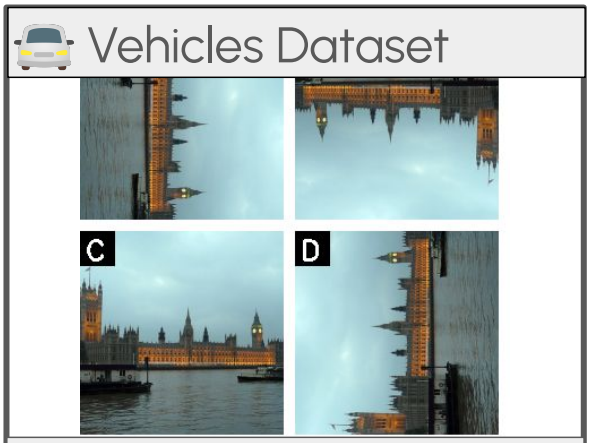
Question:
Each banana is marked with a coloured bounding box. **Which banana is closer to the camera?** Only respond with the color of the bounding box.
Answer: Red



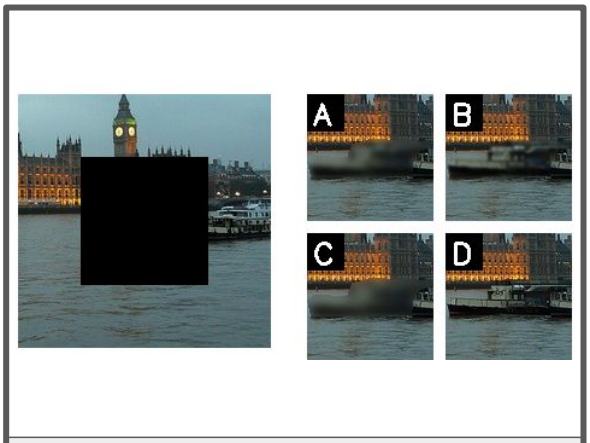
Question:
Which image variant is not corrupted? Please respond only with the letter.
Answer: A

Question:
One banana is marked with a bounding box. **Is there another banana further to the left in the image?** Please respond only with yes or no.
Answer: Yes

C.4 Vehicles Dataset Examples

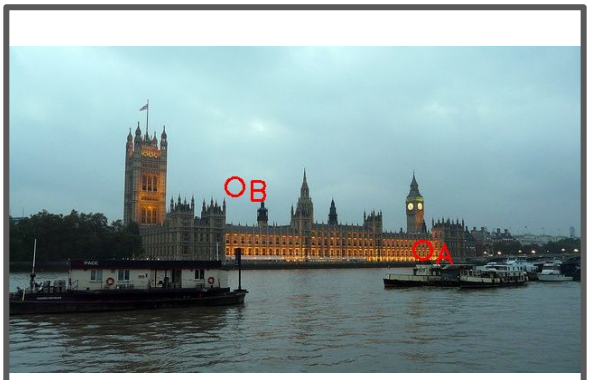


Question:
Which image variant is not rotated? Please respond only with the letter.
Answer: C

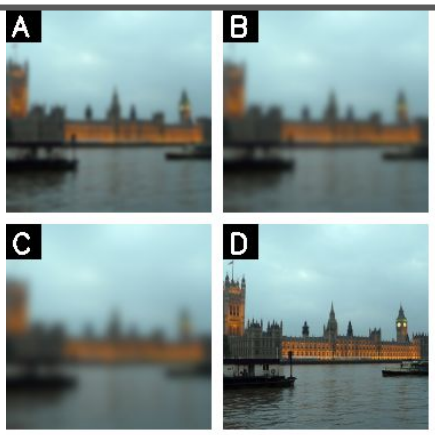


Question:
Which tile fits best in the image? Please respond only with the letter.
Answer: D

Question:
Each boat is marked with a coloured bounding box. **Which boat is further to the left in the image?** Only respond with the color of the bounding box.
Answer: Red

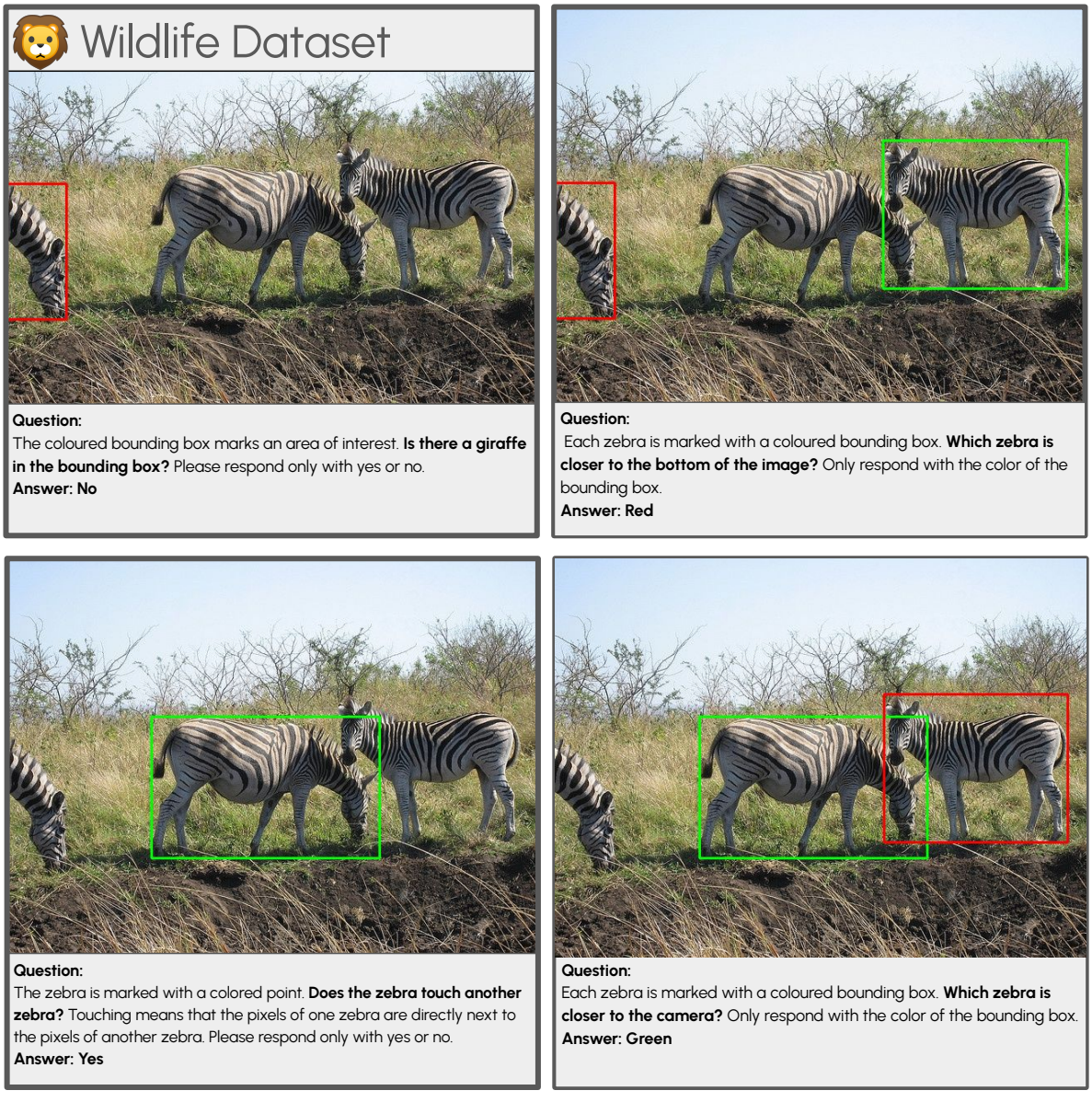


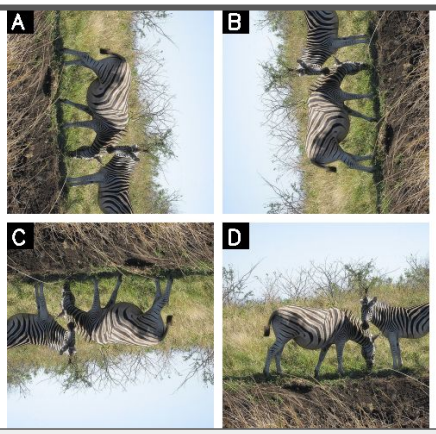
Question:
Which point is closer? Please respond only with the letter.
Answer: A



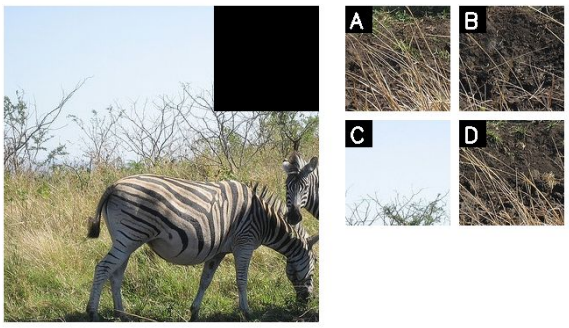
Question:
Which image variant contains the smallest amount of blur? Please respond only with the letter.
Answer: D

C.5 Wildlife Dataset Examples





Question:
Which image variant is not rotated? Please respond only with the letter.
Answer: D

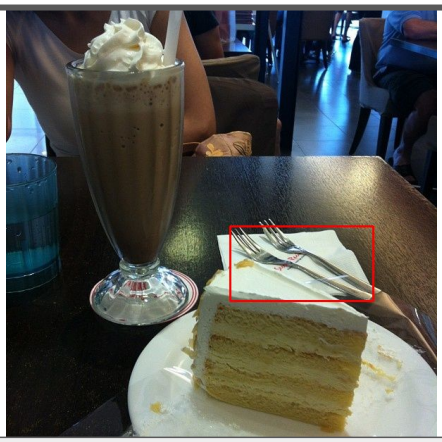


Question:
Which tile fits best in the image? Please respond only with the letter..
Answer: C

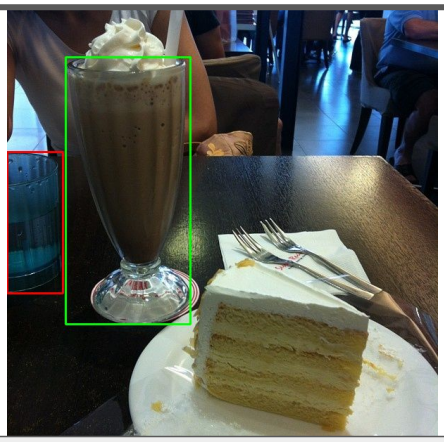
C.6 Kitchen Dataset Examples

Kitchen Dataset

Question:
How many forks are visible in the image? Please respond only with the number.
Answer: 2

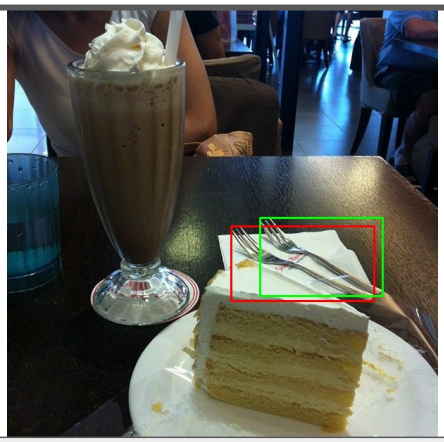


Question:
The fork is marked with a colored point. **Does the fork touch another fork?** Touching means that the pixels of one fork are directly next to the pixels of another fork. Please respond only with yes or no.
Answer: Yes

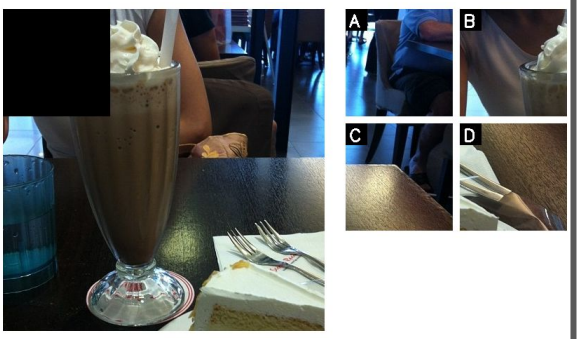


Question:
Each cup is marked with a coloured bounding box. **Which cup is closer to the camera?** Only respond with the color of the bounding box.
Answer: Green

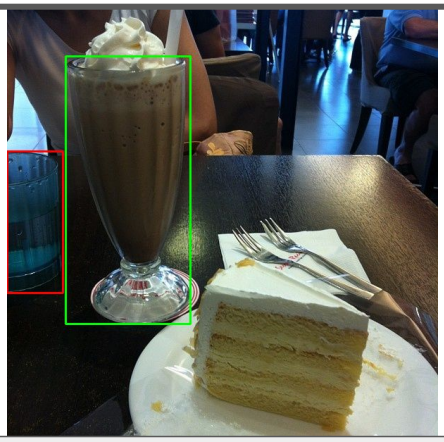
Question:
Which point is closer? Please respond only with the letter.
Answer: B



Question:
Each fork is marked with a coloured bounding box. **Which fork is closer to the bottom of the image?** Only respond with the color of the bounding box.
Answer: Red



Question:
Which tile fits best in the image? Please respond only with the letter..
Answer: B



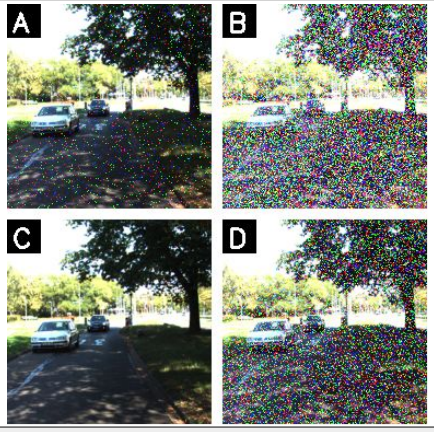
Question:
Each cup is marked with a coloured bounding box. **Which cup occupies more pixels in the image?** Only respond with the color of the bounding box.
Answer: Green

C.7 Kitti Dataset Examples

Kitti Dataset

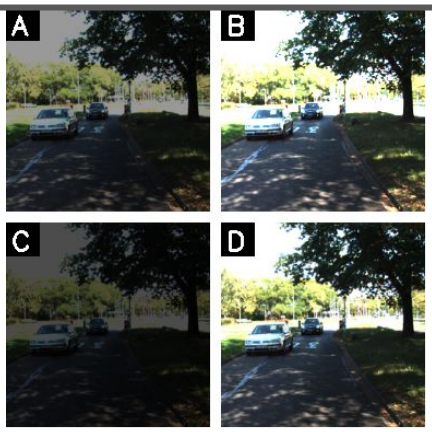
Question:
The coloured bounding box marks an area of interest. **Is there an car in the bounding box?** Please respond only with yes or no.
Answer: Yes

Question:
Which of the four tiles fits best in the image? Please respond only with the letter.
Answer: B



Question:
Which image variant is not corrupted? Please respond only with the letter.
Answer: C

Question:
Each car is marked with a coloured bounding box. **Which car is further to the left in the image?** Only respond with the color of the bounding box.
Answer: Green



Question:
Which image variant is the second brightest version of the image?
Please respond only with the letter.
Answer: D

Question:
Which point is brighter? Please respond only with the letter.
Answer: B

C.8 API Errors and Safety Settings of the Gemini API

When conducting experiments with the Gemini API, we had to modify the default safety settings to accommodate our use case, which was already surprising. While the text safety settings could be adjusted, the image safety settings were locked and required access through a higher-tier customer account. This limitation was particularly notable given that our experiments exclusively involved standard computer vision images. Consequently, this restriction resulted in 2% of our tasks remaining unanswered. In Figure A.6 and Figure A.7 we display some tasks that triggered the safety settings.

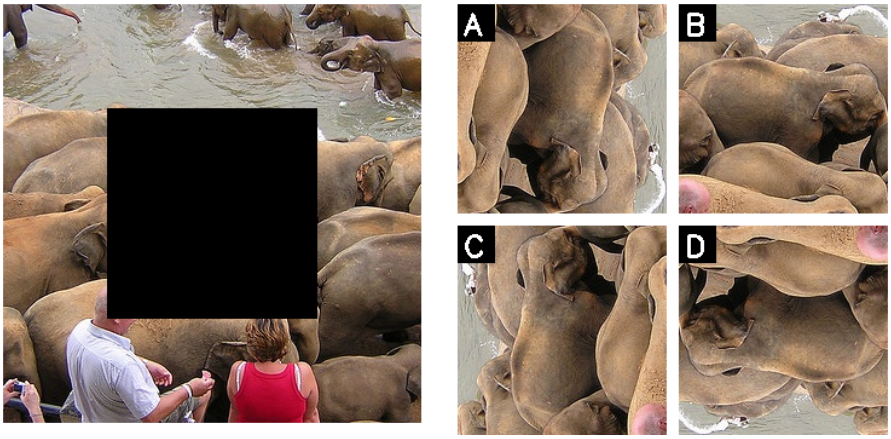


Figure A.6: **Safety systems in vision-language models can be triggered by benign inputs.** Example showing an image that activated Google Gemini's content filtering mechanisms despite containing no harmful content.

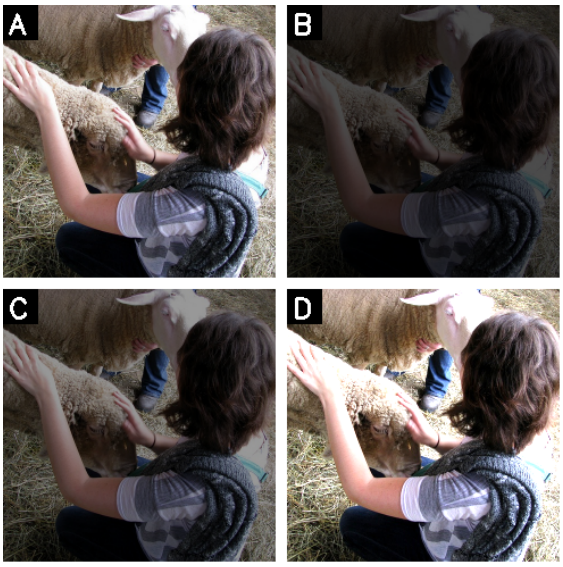


Figure A.7: **Safety systems in vision-language models can be triggered by benign inputs.** Example showing an image that activated Google Gemini's content filtering mechanisms despite containing no harmful content.

D Accuracy Analyses

D.1 Accuracy%(t) Thresholds for All Datasets

Here we present the Accuracy%(t) metric at thresholds 0.4, 0.5, 0.55, 0.6, 0.65, 0.70, 0.75, and 0.80. The top 10 models for each dataset are displayed as bar plots in Figure A.8. To increase interpretability of single models across datasets, Figure A.9 displays all 22 models as line plots.

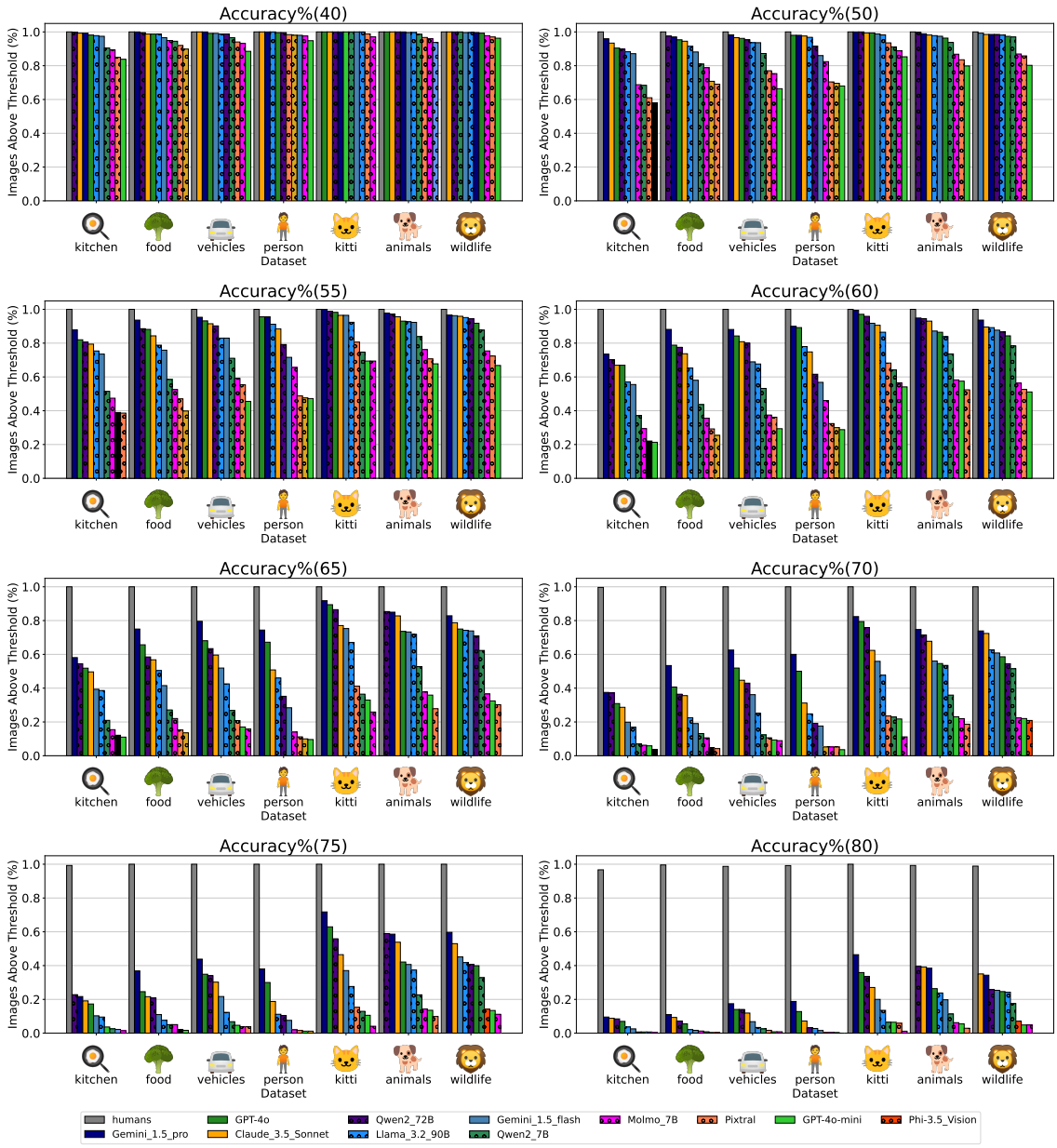


Figure A.8: **Accuracy% (t) ↑thresholds for the top 10 models across all datasets**. Humans (solid grey bar) consistently achieve near-perfect scores of 1 across all displayed thresholds and datasets. Performance varies significantly across imaging domains, with dotted lines representing open-source models and solid lines indicating closed-source models.

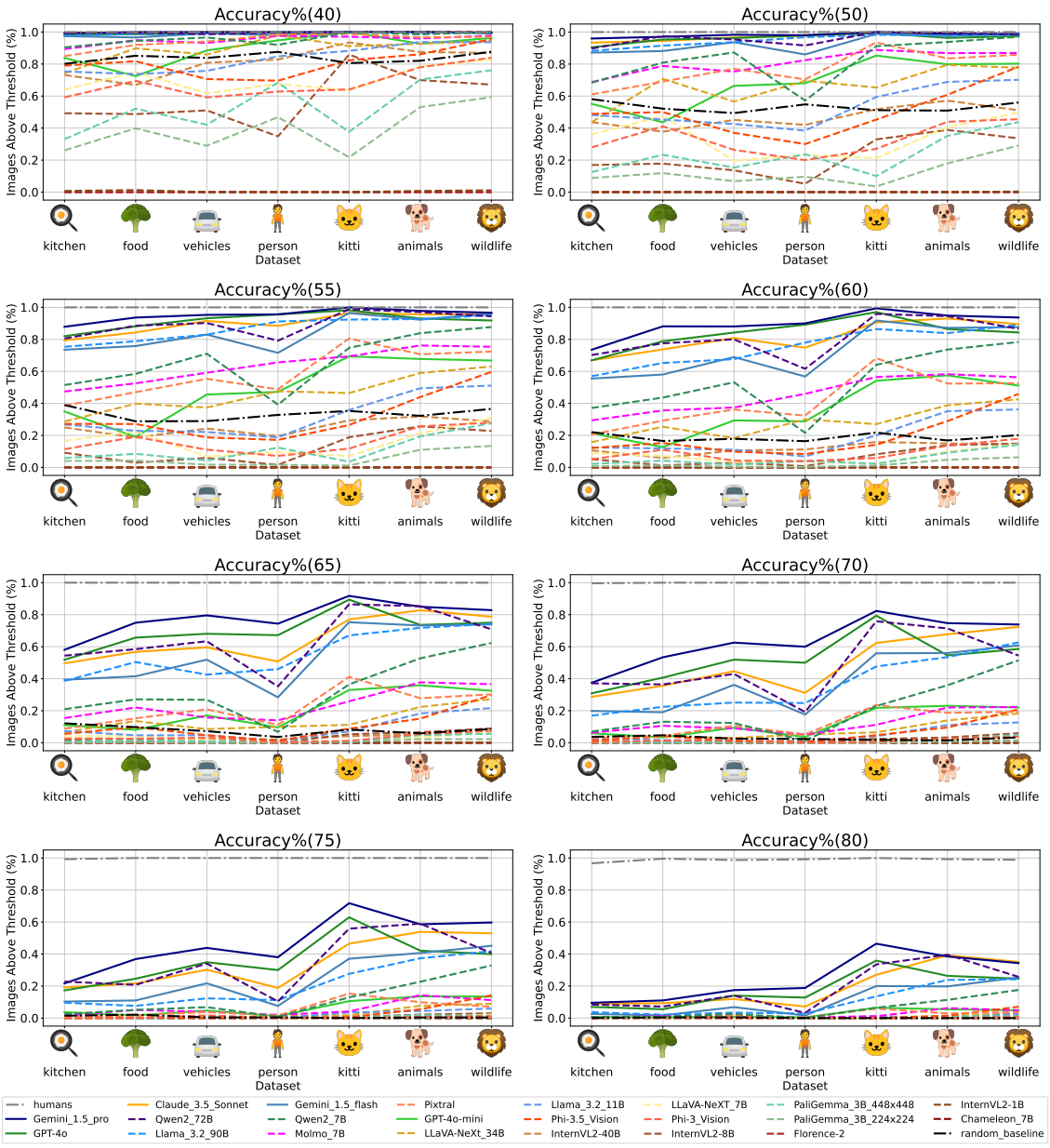


Figure A.9: **Accuracy%(t) ↑thresholds for all models for each dataset**. Displayed as line plots, given the large number of models. Humans (dashed-dotted grey line) consistently achieve near-perfect scores of 1 across all displayed thresholds and datasets. Performance varies significantly across imaging domains, with dotted lines representing open-source models and solid lines indicating closed-source models.

D.2 Accuracies Across Datasets

We display the regular Accuracy scores for all models per dataset in Table A.4 and per task in Table A.6.

Model	Overall	wildlife	animals	kitti	person	vehicles	food	kitchen
Humans	94.9	95.7	95.6	95.2	96.6	93.2	94.2	93.5
Gemini Pro	74.8	77.6	78.0	79.5	74.1	74.9	71.2	68.2
Claude_3.5_Sonnet	71.0	76.9	76.6	74.0	68.0	70.7	66.0	64.9
GPT-4o	70.8	73.5	73.0	76.7	70.1	70.8	66.8	64.3
Qwen2_72B	70.3	72.7	76.5	75.5	63.9	70.3	67.3	65.6
Llama_3.2_90B	67.5	73.5	72.1	70.8	65.0	66.2	63.3	61.7
Gemini Flash	67.5	74.5	71.7	71.1	62.8	67.9	62.3	62.2
Qwen2_7B	62.1	70.6	67.6	64.3	54.3	62.0	58.6	57.0
Molmo_7B	61.0	65.2	65.3	62.4	59.4	59.1	59.6	56.2
Pixtral	59.2	64.1	62.5	64.3	57.0	59.2	55.1	52.5
GPT-4o-mini	57.1	62.7	62.1	62.0	55.7	56.3	48.4	52.8
LLaVA-NeXt_34B	55.4	60.6	59.5	56.0	56.0	52.7	54.5	48.4
InternVL2-40B	52.7	55.7	55.2	56.7	51.1	53.4	47.0	49.8
Phi-3.5_Vision	50.7	61.1	53.9	49.2	45.0	46.8	50.3	48.4
Llama_3.2_11B	49.8	55.1	54.3	50.9	47.7	47.2	46.3	47.2
LLaVA-NeXT_7B	45.6	51.1	47.9	42.9	43.2	42.5	47.4	44.2
Phi-3_Vision	45.3	51.0	48.9	43.4	41.4	42.1	46.9	43.4
InternVL2-8B	43.9	48.0	48.9	49.9	38.1	42.1	39.9	40.0
LLaVA-v1.5_7B	40.6	46.3	39.1	39.8	41.6	37.1	41.5	38.9
PaliGemma_3B_448x448	40.6	46.5	44.5	35.9	42.2	37.7	41.2	35.8
PaliGemma_3B_224x224	36.3	41.2	39.8	32.2	38.2	33.8	36.5	32.6
InternVL2-1B	18.3	20.9	17.5	20.9	16.4	15.8	20.8	15.5
Florence-2	16.2	19.3	17.2	16.3	19.0	14.1	16.9	10.7
Chameleon_7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table A.4: **Model Accuracies across different datasets.** Performance varies greatly between domain datasets, highlighting the need for in-domain validation. For each column, the top 10 models are highlighted: 1st place (Gold) 2nd place (Silver) 3rd place (Bronze) 4th place 5th place 6th place 7th place 8th place 9th place 10th place. The 'Overall' column represents the mean accuracy across all datasets.

D.3 Accuracy%(t) Curves for All Models

Here we present the Accuracy%(t) curves for all models and datasets. The Accuracy%(t) metric represents the percentage of images for which at least a specified proportion of questions are correctly answered. The thresholds for each curve are [0.2, 0.3, 0.4, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0]. First we display the top 10 models per dataset in Figure A.10, for all 22 models in Figure A.11, and Area under the Accuracy%(t) Curves in Table A.5.

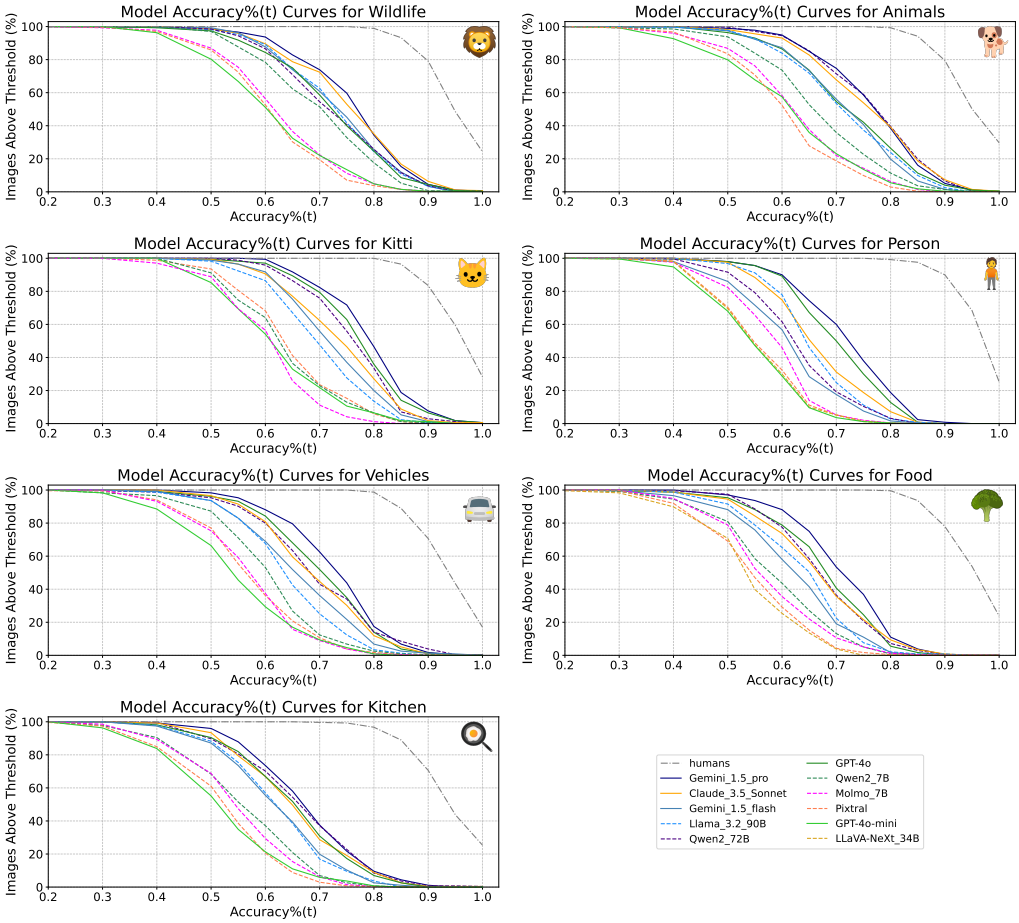


Figure A.10: **Accuracy% (t) curves for the top 10 models across each dataset,** with a maximum score of 1. Humans (dashed-dotted grey line) consistently achieve the highest performance. Dashed lines indicate open-source models, while solid lines represent closed-source models. The area under the Accuracy% (t) curves, detailed in Table A.5, highlights significant variations in model rankings across domain-specific datasets for the same tasks.

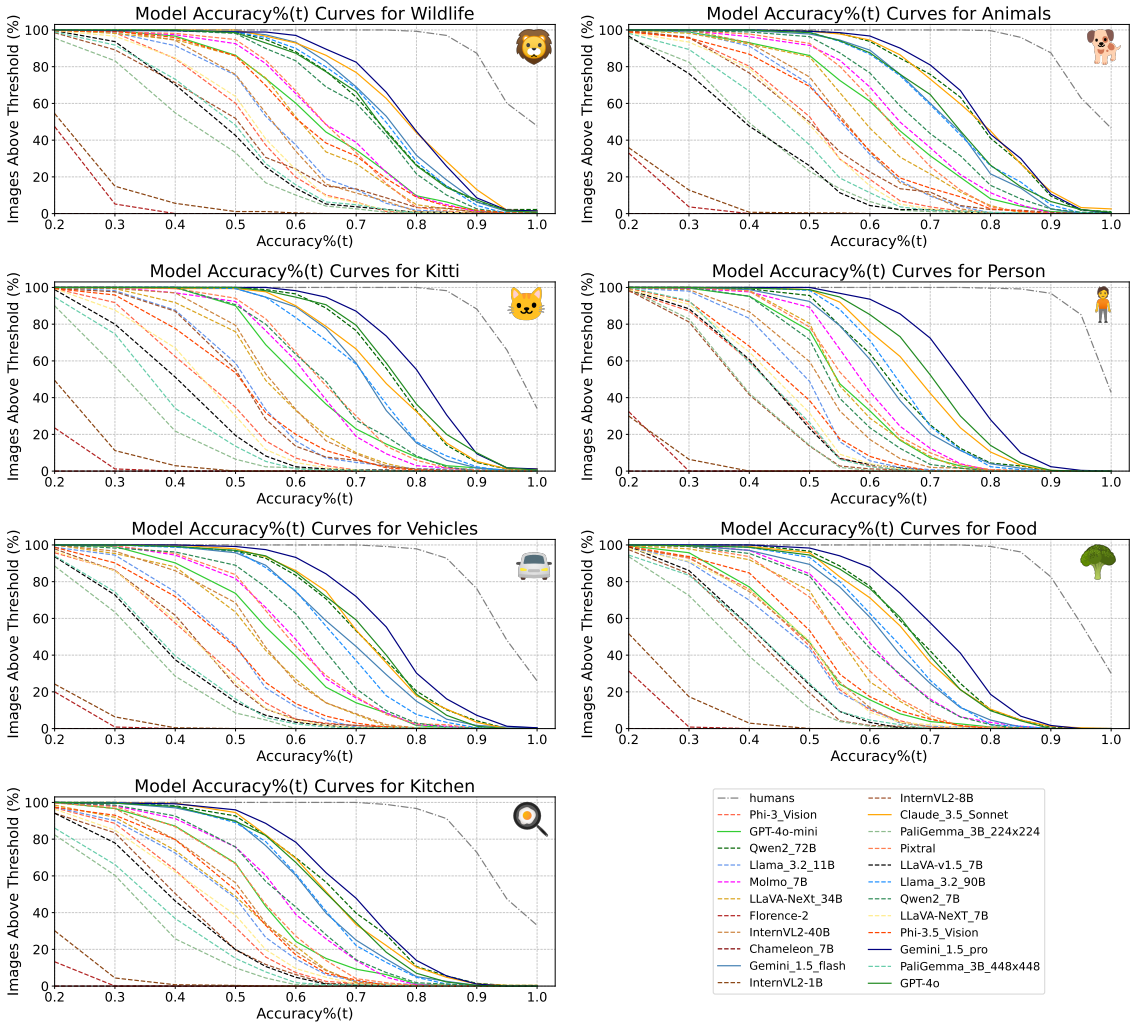


Figure A.11: **Accuracy% (t) curves for all models across each dataset,** with a maximum score of 1. Humans (dashed-dotted grey line) consistently achieve the highest performance. Dashed lines indicate open-source models, while solid lines represent closed-source models. The area under the Accuracy% (t) curves, detailed in Table A.5, highlights significant variations in model rankings across domain-specific datasets for the same tasks.

Model	Overall	animals	food	kitchen	kitti	person	vehicles	wildlife
Humans	75.42	75.97	75.01	73.68	75.97	77.61	73.85	75.88
Gemini Pro	55.18	58.42	51.44	48.41	60.06	54.48	55.30	58.12
Claude_3.5_Sonnet	51.50	57.47	46.29	45.83	54.16	47.93	51.13	57.67
GPT-4o	50.88	53.27	46.83	44.60	57.01	50.09	51.06	53.28
Qwen2_72B	50.73	57.28	47.63	46.48	55.97	43.57	50.76	53.42
Llama_3.2_90B	47.81	52.74	43.35	42.54	50.99	44.70	46.38	53.95
Gemini Flash	47.77	52.35	42.34	42.64	51.32	42.59	48.04	55.11
Qwen2_7B	42.26	48.32	38.62	37.43	44.26	33.93	41.97	51.31
Molmo_7B	40.98	45.26	39.52	36.84	42.04	38.95	39.04	45.21
Pixtral	39.43	43.16	35.01	33.06	44.29	36.88	39.16	44.48
GPT-4o-mini	37.55	42.84	28.57	33.38	42.13	35.59	36.56	43.74
LLaVA-NeXt_34B	35.44	39.50	34.38	28.56	35.63	35.99	32.73	41.29
InternVL2-40B	32.86	35.74	27.27	30.08	36.40	31.09	33.10	36.32
Phi-3.5_Vision	31.22	35.01	29.98	29.10	29.93	25.34	26.94	42.23
Llama_3.2_11B	30.17	35.04	26.04	27.32	31.41	27.98	27.19	36.19
random_chance	29.87	30.43	30.00	29.96	29.50	29.85	30.78	28.55
LLaVA-NeXT_7B	25.87	28.50	27.00	24.22	23.63	23.59	22.12	32.06
Phi-3_Vision	25.82	29.52	26.99	23.81	24.34	22.00	22.54	31.56
InternVL2-8B	24.69	30.51	20.42	20.62	30.56	18.46	22.46	29.80
PaliGemma_3B_448x448	21.19	25.26	21.31	16.30	17.01	22.66	17.87	27.88
LLaVA-v1.5_7B	21.18	20.23	21.59	19.47	20.16	22.20	17.45	27.15
PaliGemma_3B_224x224	17.24	20.89	16.97	13.74	13.10	18.65	14.49	22.83
InternVL2-1B	3.24	3.20	4.62	2.05	3.88	2.14	1.89	4.93
Florence-2	1.60	2.01	1.65	0.66	1.29	1.62	1.09	2.89
Chameleon_7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A.5: **Area under Accuracy%(t) Curves for seven different datasets**. Curves calculated for threshholds at [0.2, 0.3, 0.4, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0] across each dataset. For each column, the top 10 models are highlighted: 1st place (Gold) 2nd place (Silver) 3rd place (Bronze) 4th place 5th place 6th place 7th place 8th place 9th place 10th place. The 'Overall' column represents the mean area under the curve across all datasets. The substantial differences in model rankings across domain-specific datasets for identical tasks highlights the need for specific in-domain evaluation.

E Ranking Analyses

E.1 Ranking Variability Across Datasets

Scatter plots illustrating ranking diversity first present a combined view across all domains, followed by separate plots for each domain. These visualizations reveal the variations in rankings both globally and within datasets.

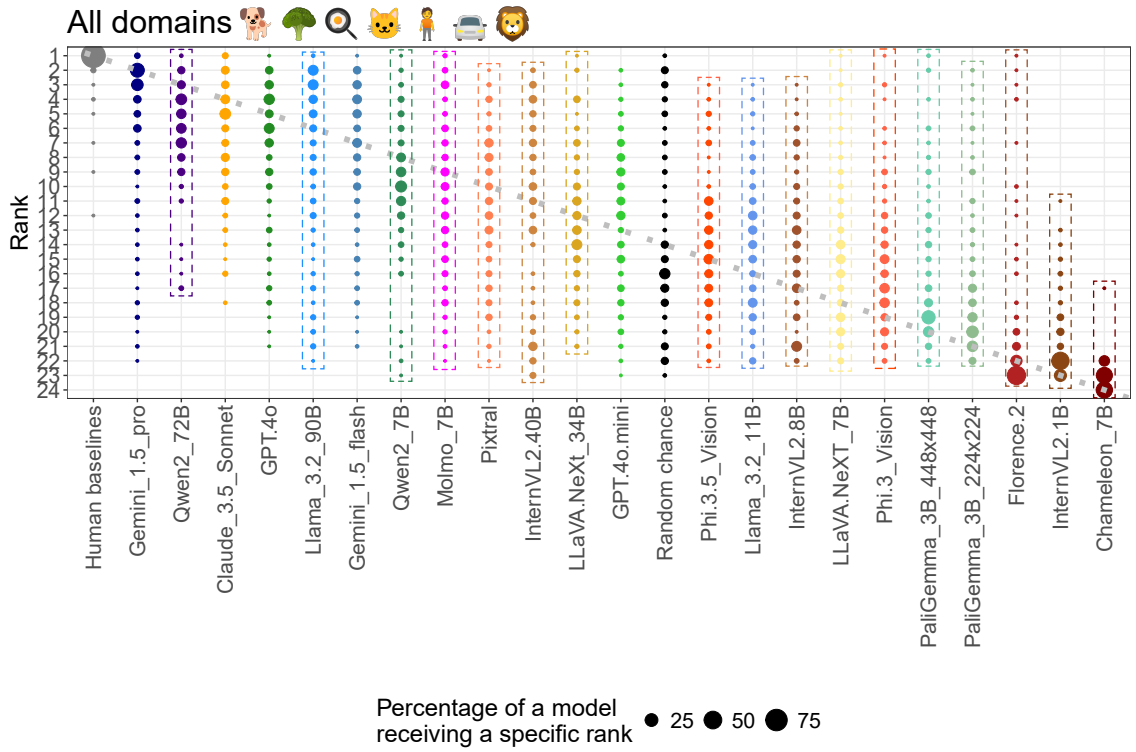


Figure A.12: **Ranking variation across models for all datasets combined.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

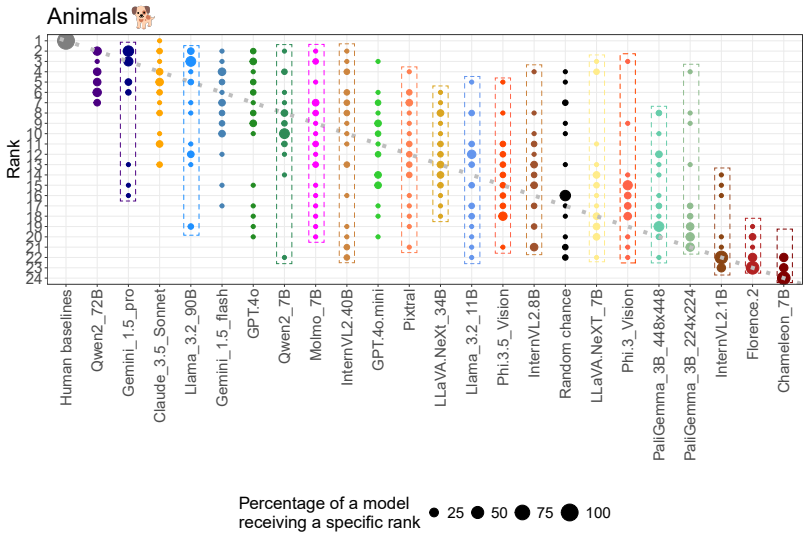


Figure A.13: **Ranking variation across models for the animals dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

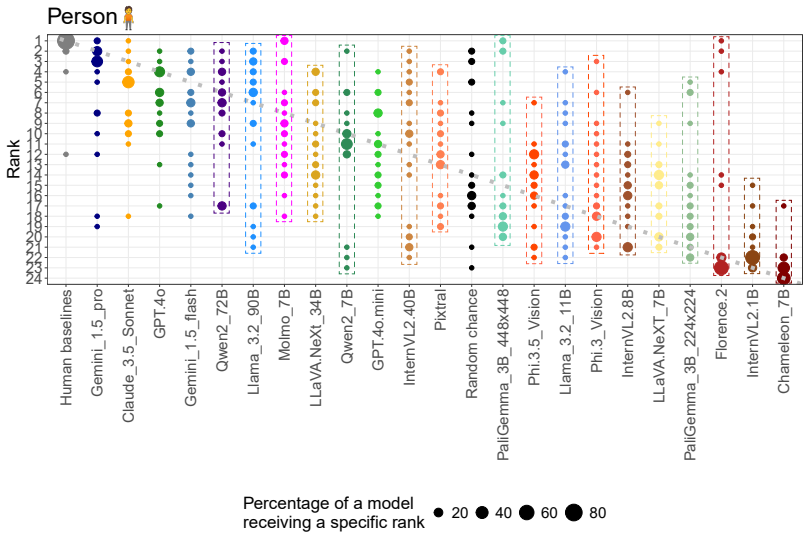


Figure A.14: **Ranking variation across models for the person dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

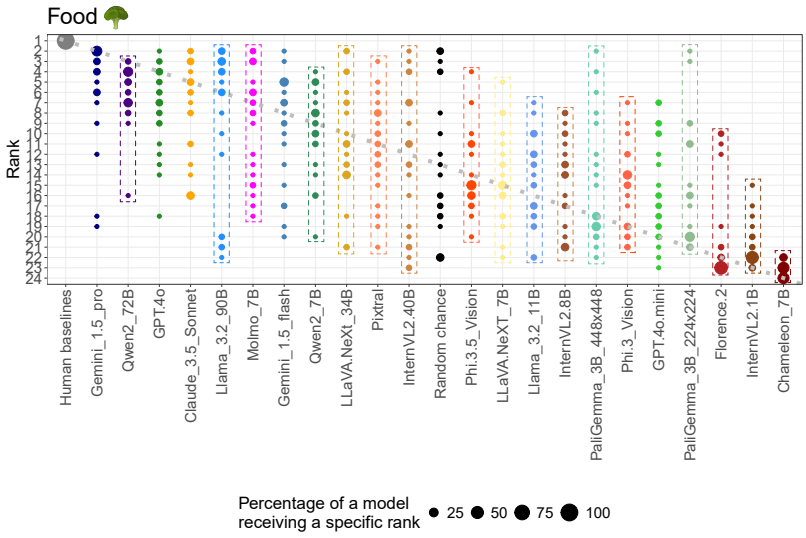


Figure A.15: **Ranking variation across models for the food dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

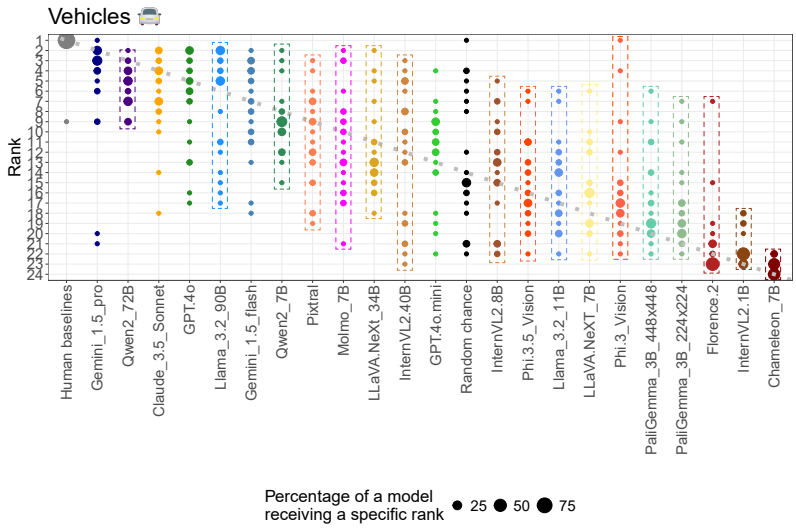


Figure A.16: **Ranking variation across models for the vehicles dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

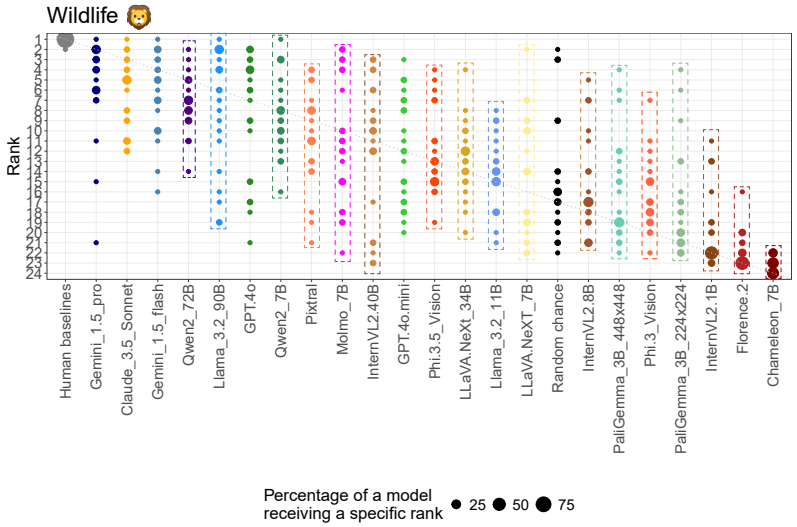


Figure A.17: **Ranking variation across models for the wildlife dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

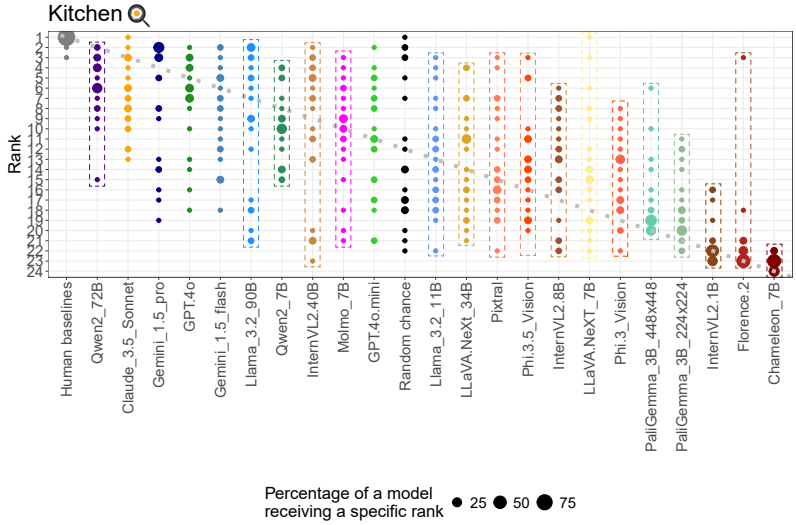


Figure A.18: **Ranking variation across models for the kitchen dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

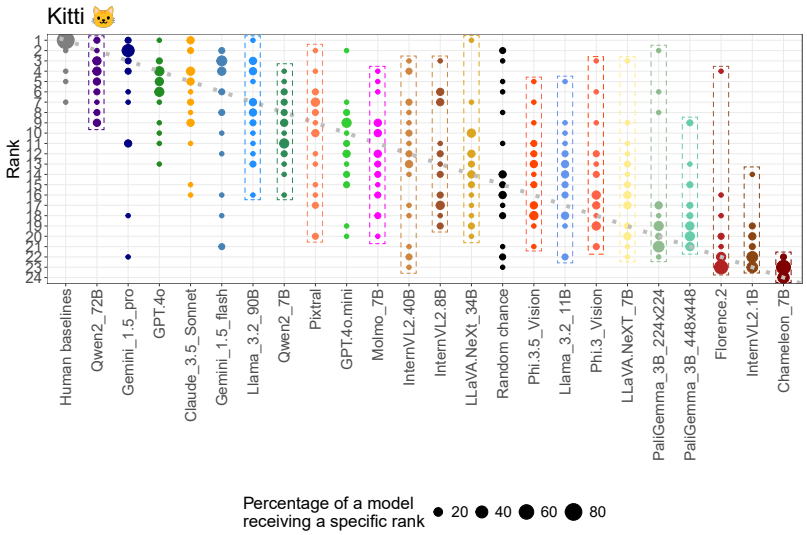


Figure A.19: **Ranking variation across models for the kitti dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

E.2 Ranking Task Comparison Between Models And Humans

Task ranking differs between models and human raters. The plots shows the difficulty of tasks based on aggregated model scores/human scores (1 = hardest task, 22 = easiest task). The radius of the blob indicates how often a task was assigned a difficulty rank when considering all seven domains and all models (n = 5 for closed models; n = 16 for open models; n = 21 for all models; n = 1 for humans as majority vote over several raters). The larger the plot, the higher the percentage it achieved a specific rank. The hardest tasks on average across domains are (1) T7.2 "Jigsaw Puzzle Completion", (2), T1.2 "Object Counting", (3), T7.1 "Rotated Jigsaw Puzzle Completion", and (4) T5.2 "Second Brightest Image Selection". The easiest task on average was T1.3 "Additional Object Presence Detection". We display aggregated all models in Figure A.20, human baselines in Figure A.21, all closed-source models in Figure A.22, and all open-source models in Figure A.23.

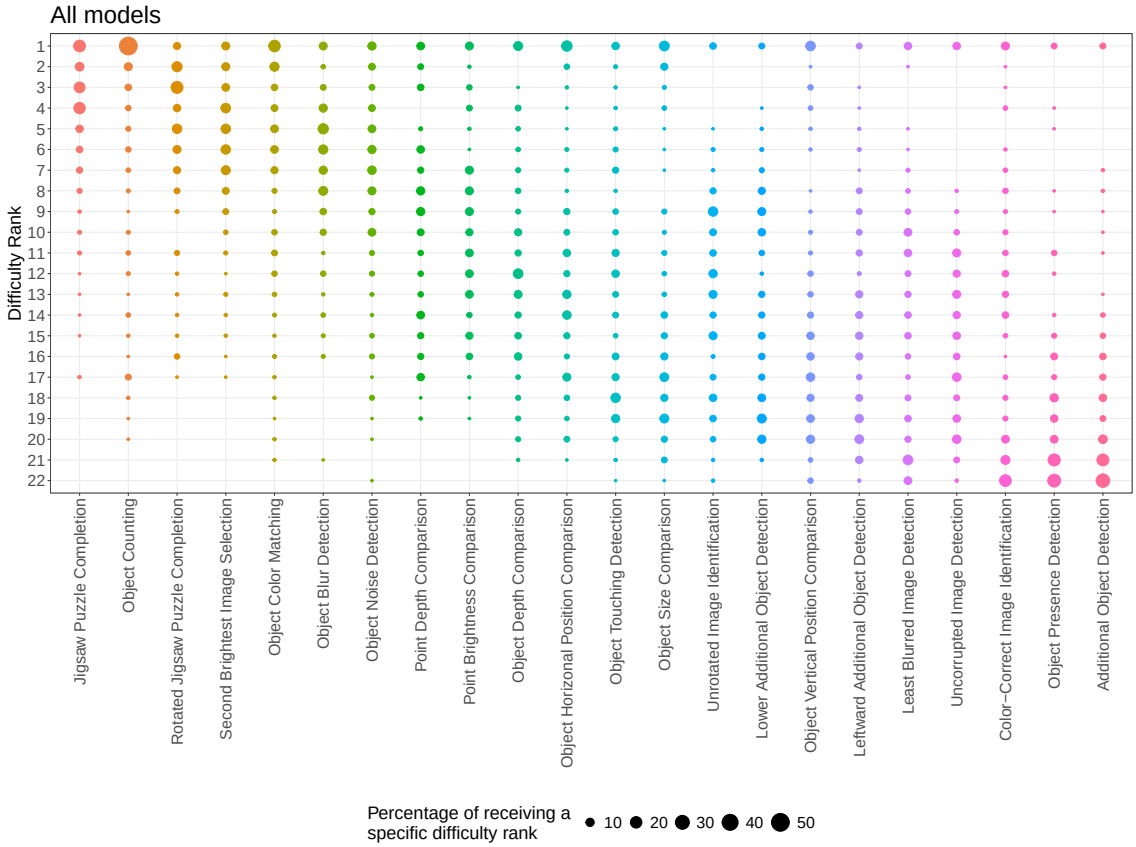


Figure A.20: **All tested vision-language models, both open and closed-source, exhibit consistent patterns in task difficulty.** Aggregated ranking of tasks from easiest to most challenging, revealing systematic strengths and limitations shared across the complete set of evaluated models regardless of their accessibility.

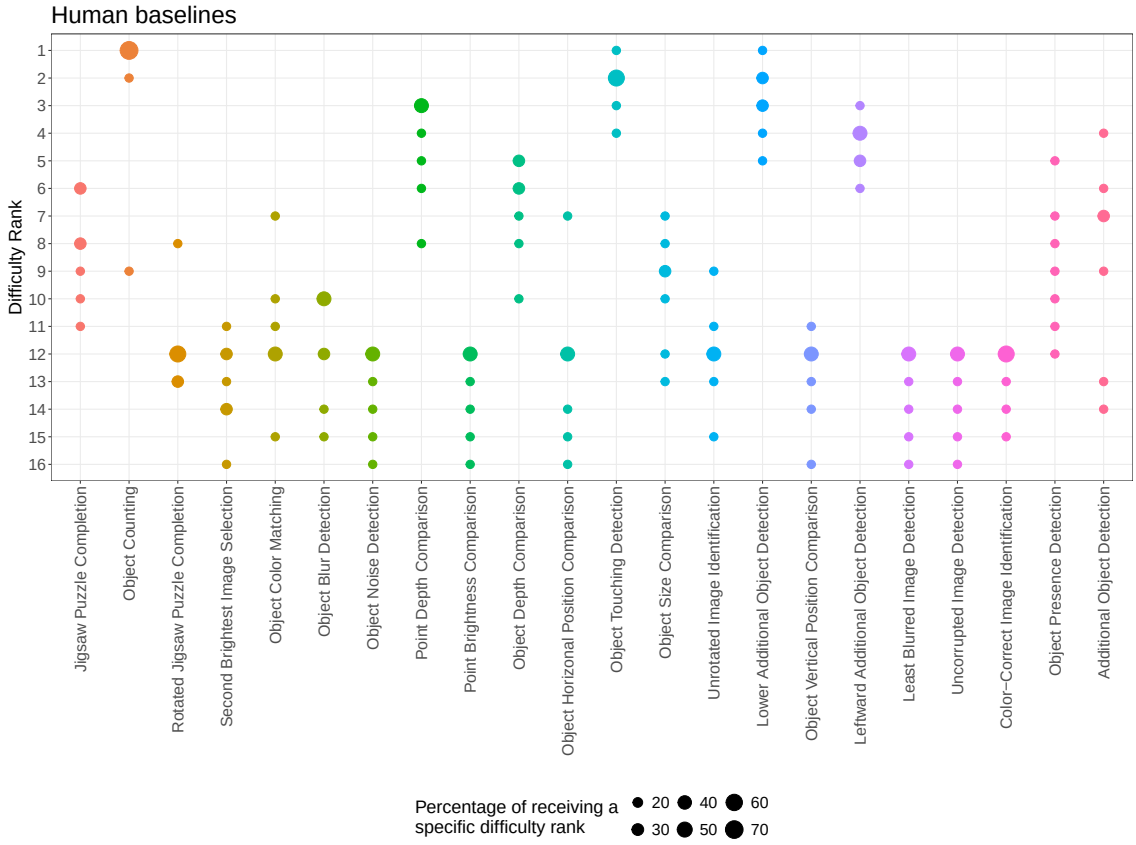


Figure A.21: **Human performance reveals distinct patterns of task difficulty compared to models.** Ranking of vision tasks from easiest to most challenging based on human evaluator performance, providing a baseline for understanding natural visual capabilities.

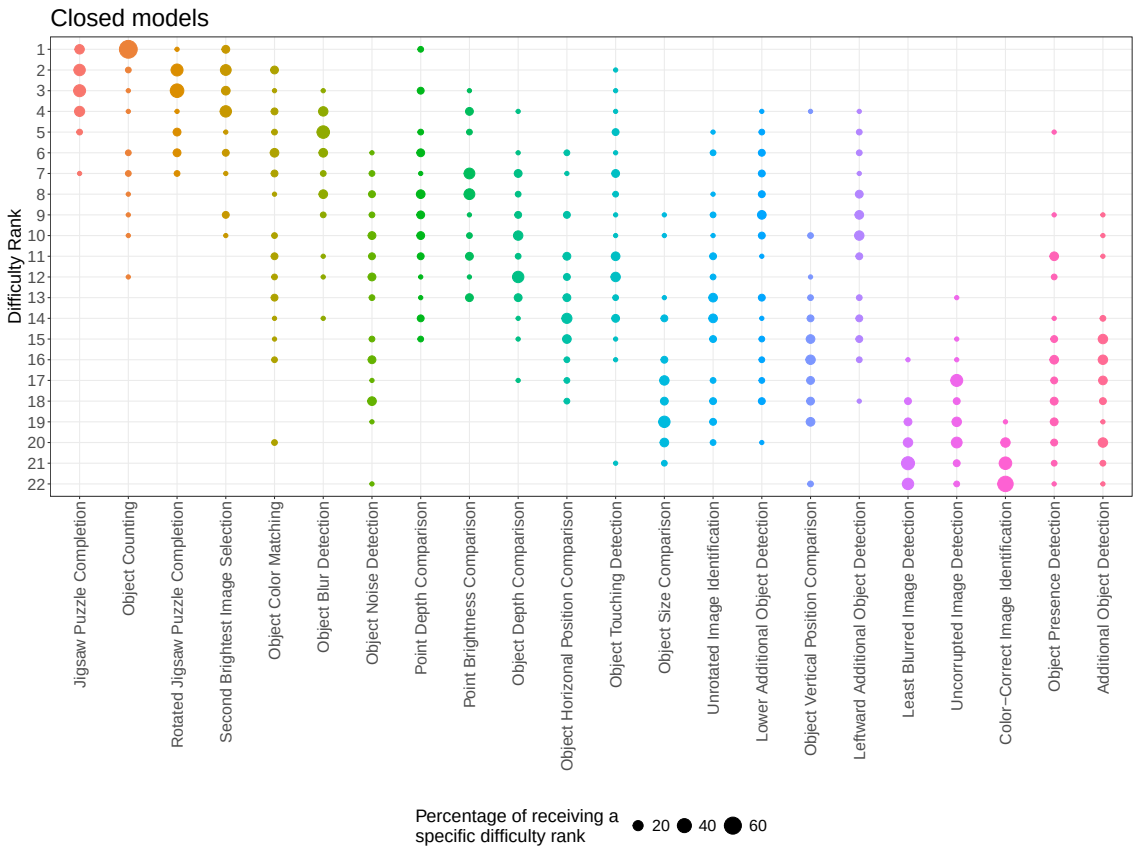


Figure A.22: **Closed-sourced vision-language models demonstrate shared patterns in task difficulty despite architectural differences.** Aggregated ranking of tasks from easiest to most challenging, revealing systematic strengths and limitations common across all evaluated closed-source models.

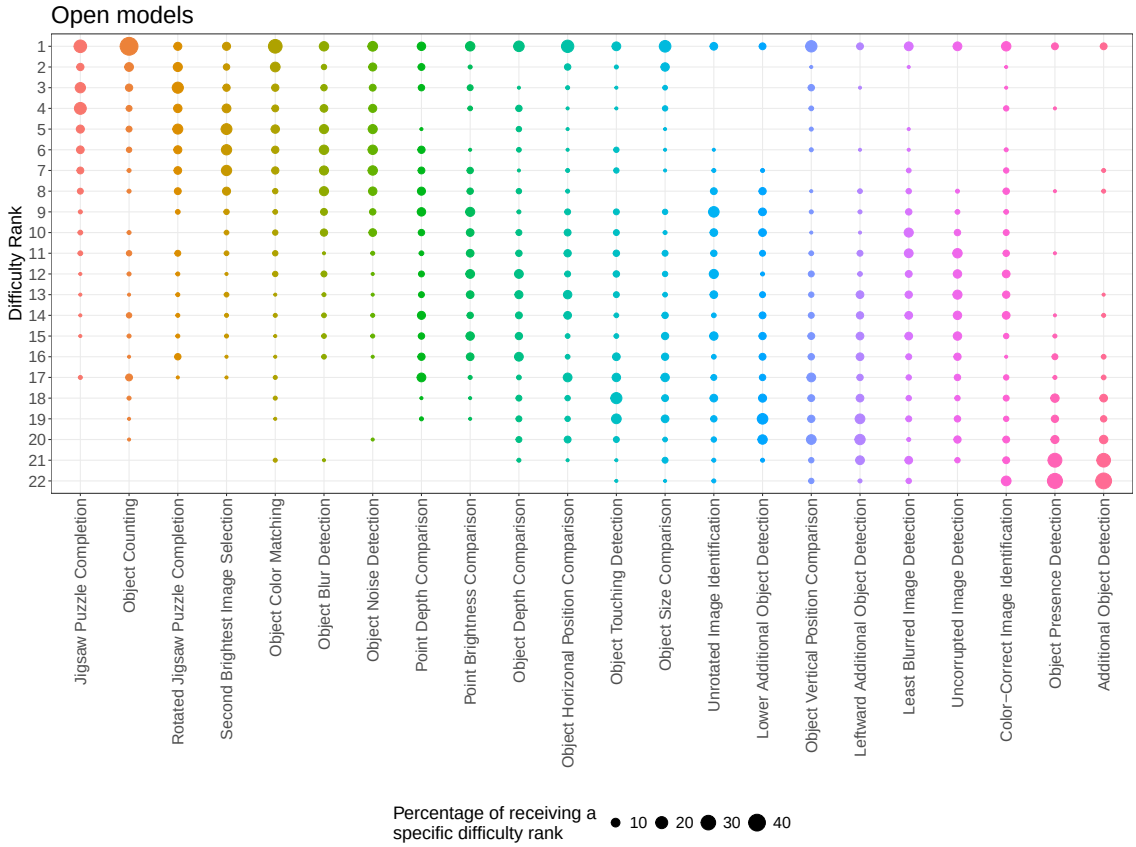


Figure A.23: **Publicly available vision-language models exhibit consistent patterns in task difficulty across architectures.** Aggregated ranking of tasks from easiest to most challenging, revealing systematic strengths and limitations shared across all evaluated open-source models.

F Task Structure and Relationships

F.1 Task Diversity: Model Accuracies Across Tasks

Here, we present the performance of models across all tasks, using Accuracy (see Table A.6). Spider plots for closed-source, open-source, and varying model sizes are displayed in Figure A.24.

Model	Rank	T1.1	T1.2	T1.3	T2.1	T2.2	T2.3	T2.4	T3.1	T3.2	T3.3	T3.4
Humans		96.88	68.31	95.24	99.51	100.00	100.00	100.00	97.21	99.83	99.90	89.74
Gemini_1.5_pro	1	78.23	33.68	92.32	75.09	93.33	97.11	98.15	96.22	75.95	92.18	74.85
GPT-4o	2	83.45	30.35	87.17	50.45	82.36	97.62	87.92	87.07	74.99	85.63	70.45
Claude_3.5_Sonnet	3	88.98	25.70	88.46	50.46	63.32	99.83	96.37	89.89	74.08	85.23	76.66
Qwen2_72B	4	92.10	34.11	91.18	45.76	62.40	96.39	85.35	86.97	71.26	86.96	82.08
Llama_3.2_90B	5	93.82	30.11	91.46	30.21	24.82	98.53	88.38	90.75	79.31	93.00	81.21
Gemini_1.5_flash	6	84.21	31.33	87.58	54.81	63.51	94.22	89.21	94.00	77.79	89.16	55.65
Qwen2_7B	7	90.65	33.82	90.11	39.38	38.14	76.98	69.56	80.16	70.88	80.03	65.24
Molmo_7B	8	90.23	30.88	84.30	27.54	30.74	78.23	76.44	79.53	74.20	82.49	67.28
Pixtral	9	86.84	7.14	88.75	41.20	34.18	78.29	85.89	71.29	63.09	75.85	72.88
GPT-4o-mini	10	79.86	23.35	67.68	43.70	60.91	80.17	72.04	62.47	65.35	59.01	54.33
LLaVA-NeXt_34B	11	88.84	31.02	90.07	30.05	31.79	57.86	61.61	74.50	64.41	79.34	76.97
Llama_3.2_11B	12	90.84	26.84	84.86	29.61	24.06	34.08	45.83	64.14	67.97	70.72	68.50
Phi-3.5_Vision	13	86.88	23.33	82.71	33.07	22.15	58.01	69.47	48.38	49.98	51.05	61.35
InternVL2-40B	14	63.66	29.59	55.06	62.62	57.09	60.56	66.63	78.00	72.34	82.92	60.35
LLaVA-NeXT_7B	15	87.08	0.00	86.71	25.41	26.43	33.54	36.81	53.94	57.34	63.21	73.26
Phi-3_Vision	16	90.54	24.99	82.70	28.65	23.84	50.31	75.56	36.55	51.77	52.41	55.61
InternVL2-8B	17	86.78	24.47	85.39	28.01	31.17	46.47	49.34	7.09	3.66	6.19	68.01
PaliGemma_3B_448x448	18	79.68	31.31	89.50	26.39	23.78	37.00	28.39	17.74	26.39	26.59	74.15
PaliGemma_3B_224x224	19	79.85	26.06	89.28	25.35	22.89	23.61	25.53	12.55	10.39	13.15	73.53
InternVL2-1B	20	31.55	0.00	47.84	24.17	24.58	15.81	25.16	0.00	0.00	0.00	34.83
Florence-2	21	50.98	26.96	77.61	0.00	0.00	0.11	0.11	0.08	0.17	0.80	64.93
Chameleon_7B	22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Model	Rank	T3.5	T4.1	T5.1	T5.2	T5.3	T5.4	T6.1	T6.2	T7.1	T7.2	T8.1
Humans		82.57	75.34	98.56	99.86	99.82	100.00	93.78	90.72	99.04	96.56	99.30
Gemini_1.5_pro	1	69.02	70.32	81.41	23.64	99.15	66.80	78.46	68.36	53.67	47.86	79.66
GPT-4o	2	66.58	67.21	82.01	46.09	96.57	57.50	75.79	43.25	41.58	46.66	85.76
Claude_3.5_Sonnet	3	69.86	63.54	46.37	61.09	98.36	69.65	68.48	64.15	39.64	37.98	79.61
Qwen2_72B	4	68.21	68.30	52.56	45.81	98.63	66.89	74.69	64.21	37.12	32.74	79.51
Llama_3.2_90B	5	67.43	72.45	35.68	28.37	97.95	75.00	78.99	62.49	27.54	23.22	78.99
Gemini_1.5_flash	6	63.95	68.17	41.06	33.76	96.46	52.53	71.80	73.40	42.23	27.55	77.20
Qwen2_7B	7	64.52	68.53	33.40	30.22	87.46	61.72	75.51	56.72	32.73	25.15	70.42
Molmo_7B	8	58.12	60.06	28.79	35.34	69.54	67.60	78.68	53.21	28.87	26.05	84.32
Pixtral	9	63.37	70.13	28.00	35.47	94.85	38.95	63.01	32.66	30.41	29.98	81.35
GPT-4o-mini	10	72.69	57.68	48.57	34.58	81.02	51.40	49.48	55.02	33.01	29.64	59.98
LLaVA-NeXt_34B	11	65.84	67.52	35.69	24.83	39.62	51.00	72.04	52.33	28.19	26.04	53.84
Llama_3.2_11B	12	63.69	55.04	7.73	26.22	46.86	62.65	62.63	54.41	25.69	26.14	38.49
Phi-3.5_Vision	13	64.30	66.02	17.26	27.28	51.81	51.41	45.56	61.98	30.67	26.10	52.27
InternVL2-40B	14	56.75	9.18	60.92	33.40	82.67	16.54	79.59	11.92	34.83	32.43	65.90
LLaVA-NeXT_7B	15	69.50	62.59	23.37	26.02	24.81	49.25	56.65	55.59	23.77	24.97	28.64
Phi-3_Vision	16	58.98	66.22	22.42	26.04	46.52	18.66	42.42	16.59	27.91	25.88	44.22
InternVL2-8B	17	64.84	59.73	51.82	27.64	56.74	51.50	18.60	51.13	27.32	28.19	54.81
PaliGemma_3B_448x448	18	67.97	60.55	16.22	25.23	33.53	41.82	27.86	51.58	24.47	22.28	28.83
PaliGemma_3B_224x224	19	71.00	58.38	23.41	26.30	24.59	34.08	19.91	37.08	22.13	24.20	27.14
InternVL2-1B	20	25.56	14.39	24.38	10.64	24.42	12.82	0.00	5.82	23.82	22.14	24.71
Florence-2	21	57.33	49.88	0.00	0.70	0.00	0.17	0.26	3.56	4.83	0.00	0.65
Chameleon_7B	22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A.6: **Model performance across tasks displayed as regular Accuracy.** For each column, the top 10 models are highlighted: 1st place (Gold) 2nd place (Silver) 3rd place (Bronze) 4th place 5th place 6th place 7th place 8th place 9th place 10th place. Certain models exhibit weaknesses on specific tasks despite strong overall rankings. For instance, Gemini Pro underperforms on the "Second Brightest Image" task (T5.2, 23.6%, rank 19), in contrast to Claude 3.5 Sonnet (61.1%, rank 1). GPT-4o shows limitations on "Point Depth Comparison" (T6.2, 43.3%, rank 15) and "Additional Object Presence Detection" (T1.3, rank 11). Even Llama 3.2 90B—ranked 5th overall—struggles with "Object Blur Detection" (T2.3, 30.2%) and "Object Noise Detection" (T2.4, 24.8%) while excelling at spatial reasoning (T3.2, T3.3).

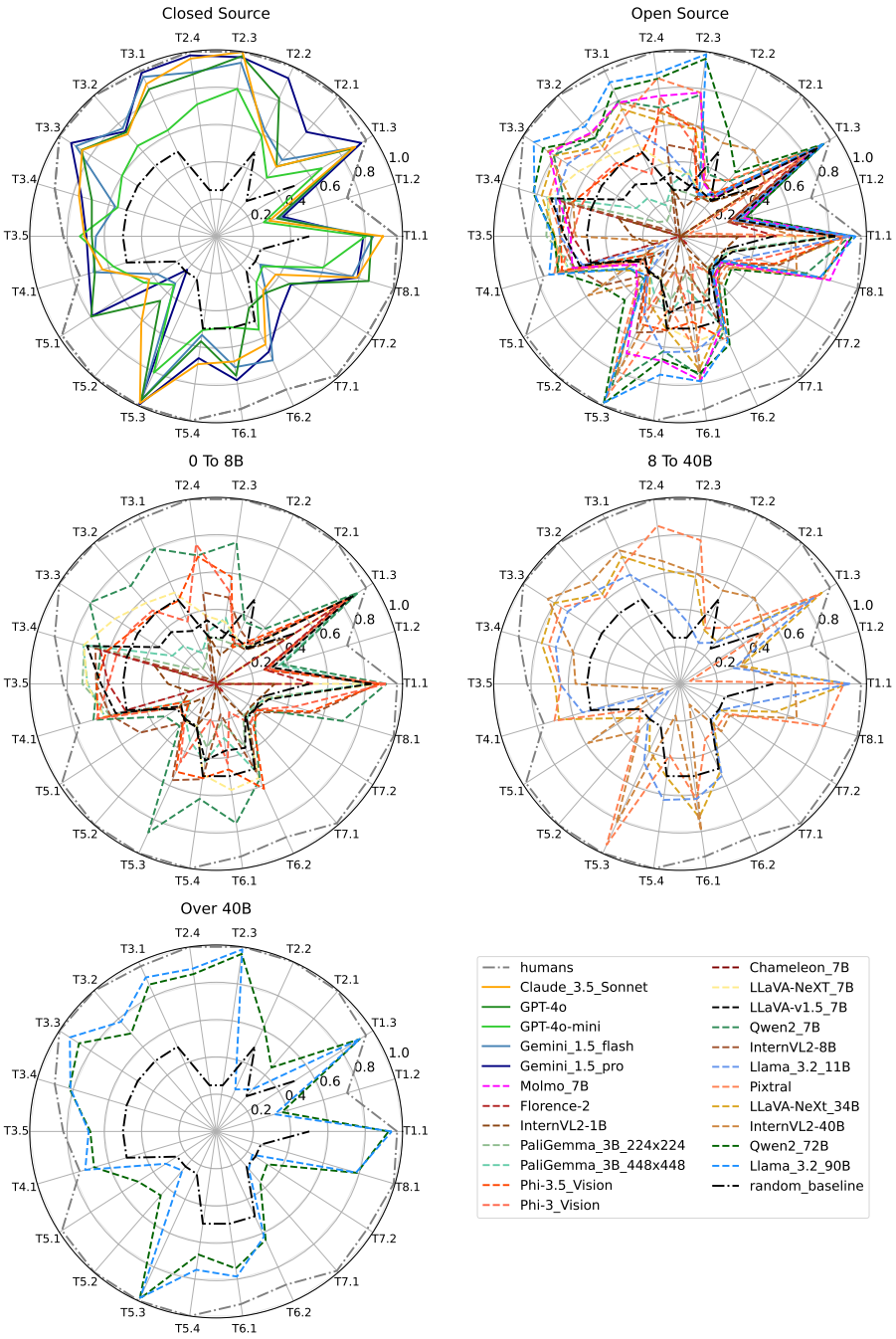


Figure A.24: **Model architecture and size significantly impact performance patterns across diverse vision tasks.** Spider plots reveal distinct performance profiles between open-source (dashed lines) and closed-source (solid lines) models across our comprehensive task framework. Each axis represents task-specific accuracy, demonstrating how different model characteristics influence capabilities.

F.2 Task Correlation

Here we present the correlation matrices for each of the seven generated domain datasets. We computed Pearson correlations between task accuracies across the five best-performing models per dataset. Most task pairs show weak correlations, with one systematic exception: tasks that modify objects—jigsaw scrambling, recoloring, and additive noise—form a tight cluster across all datasets, implying they tap into the same underlying capability. Beyond that cluster, relationships are highly dataset-specific. For example, vertical- vs. horizontal-comparison tasks correlate strongly in the Animals dataset (r = 0.973) but average only r = 0.336 elsewhere. These patterns reinforce the need for domain-specific benchmarks, as the skills probed by a task depend on the visual context.

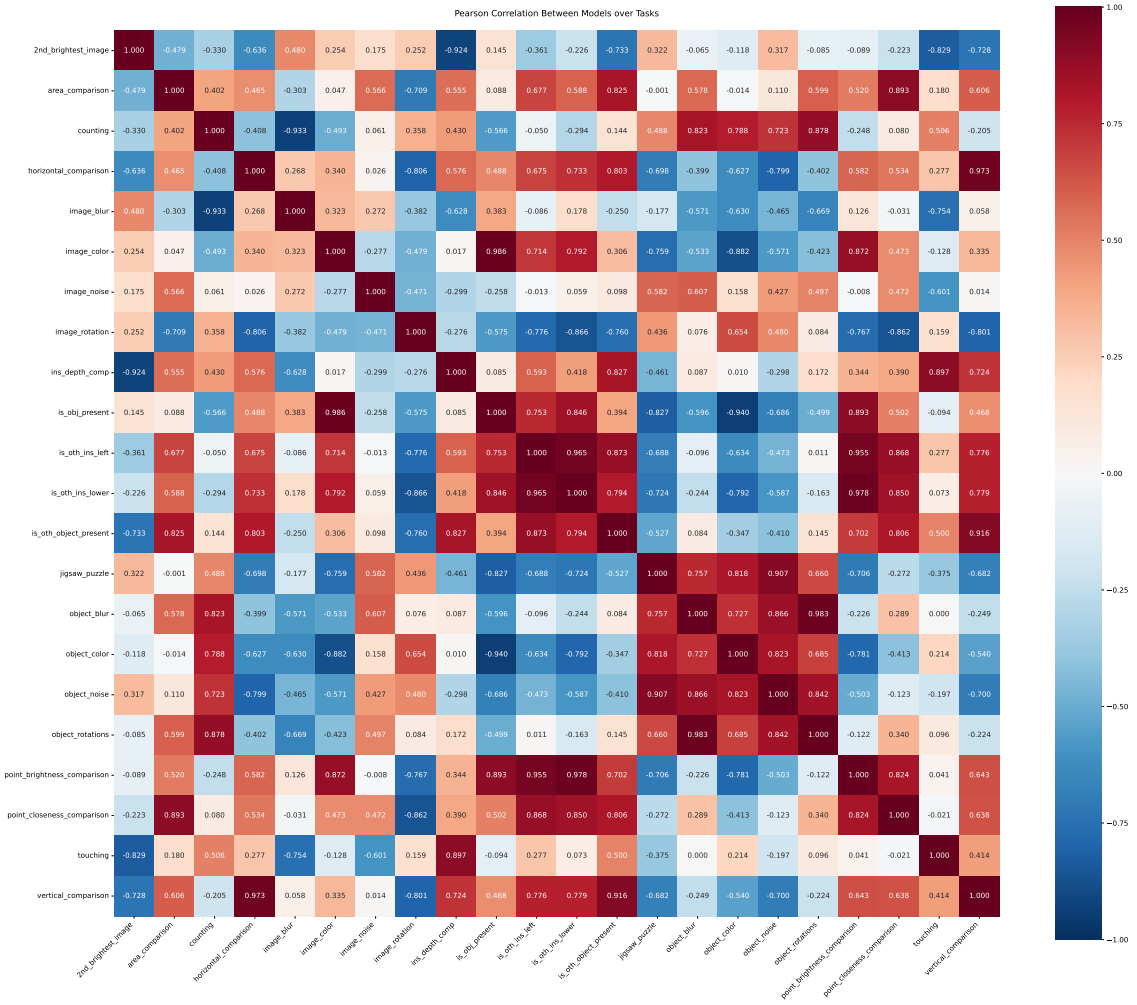


Figure A.25: **Animals Dataset: Task-performance correlations highlight shared visual capabilities.** Heatmap of Pearson correlations between task accuracies within the Animals dataset (top-5 models); coefficients range from -1 (anti-correlated) to 1 (highly correlated).

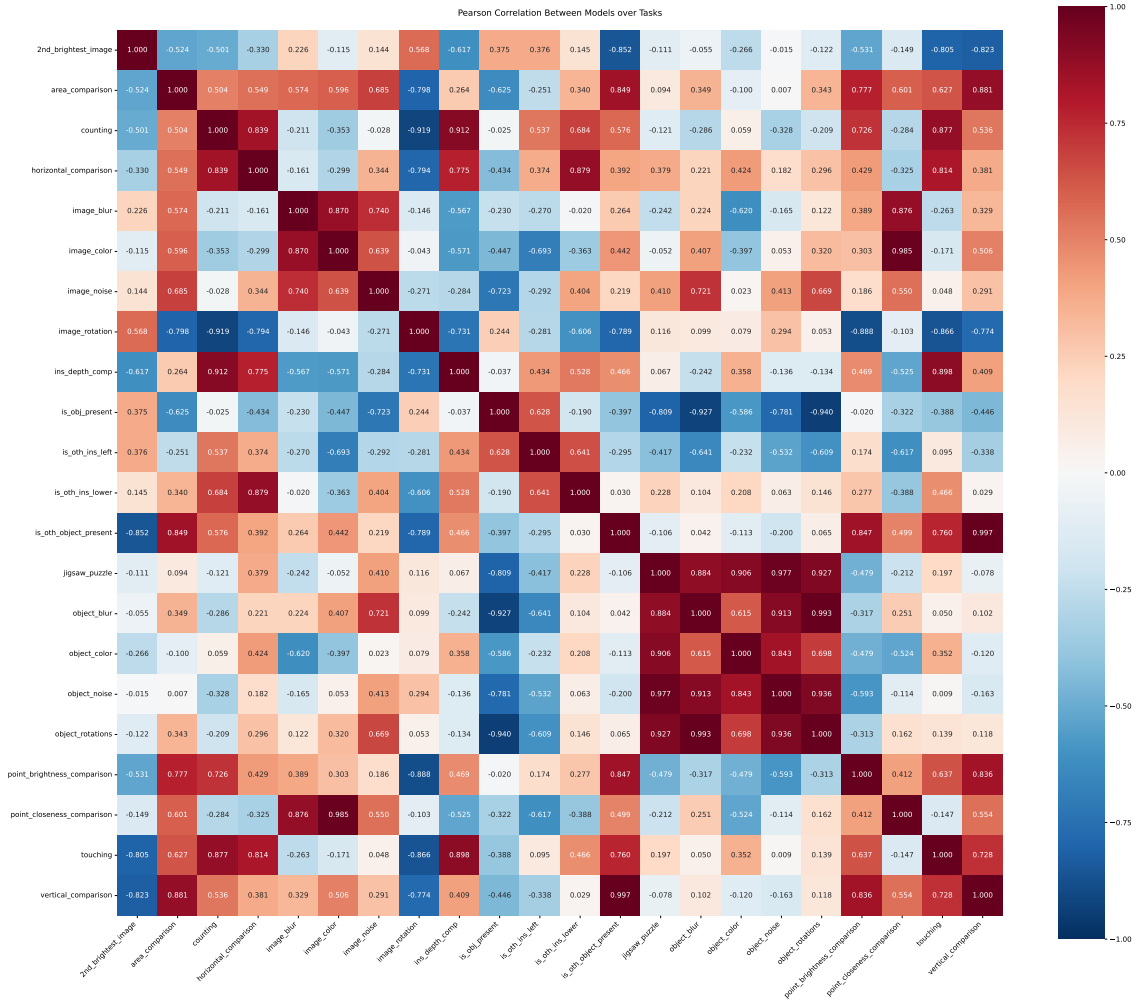


Figure A.26: **Person Dataset: Task-performance correlations highlight shared visual capabilities.** Heatmap of Pearson correlations between task accuracies within the Animals dataset (top-5 models); coefficients range from -1 (anti-correlated) to 1 (highly correlated).

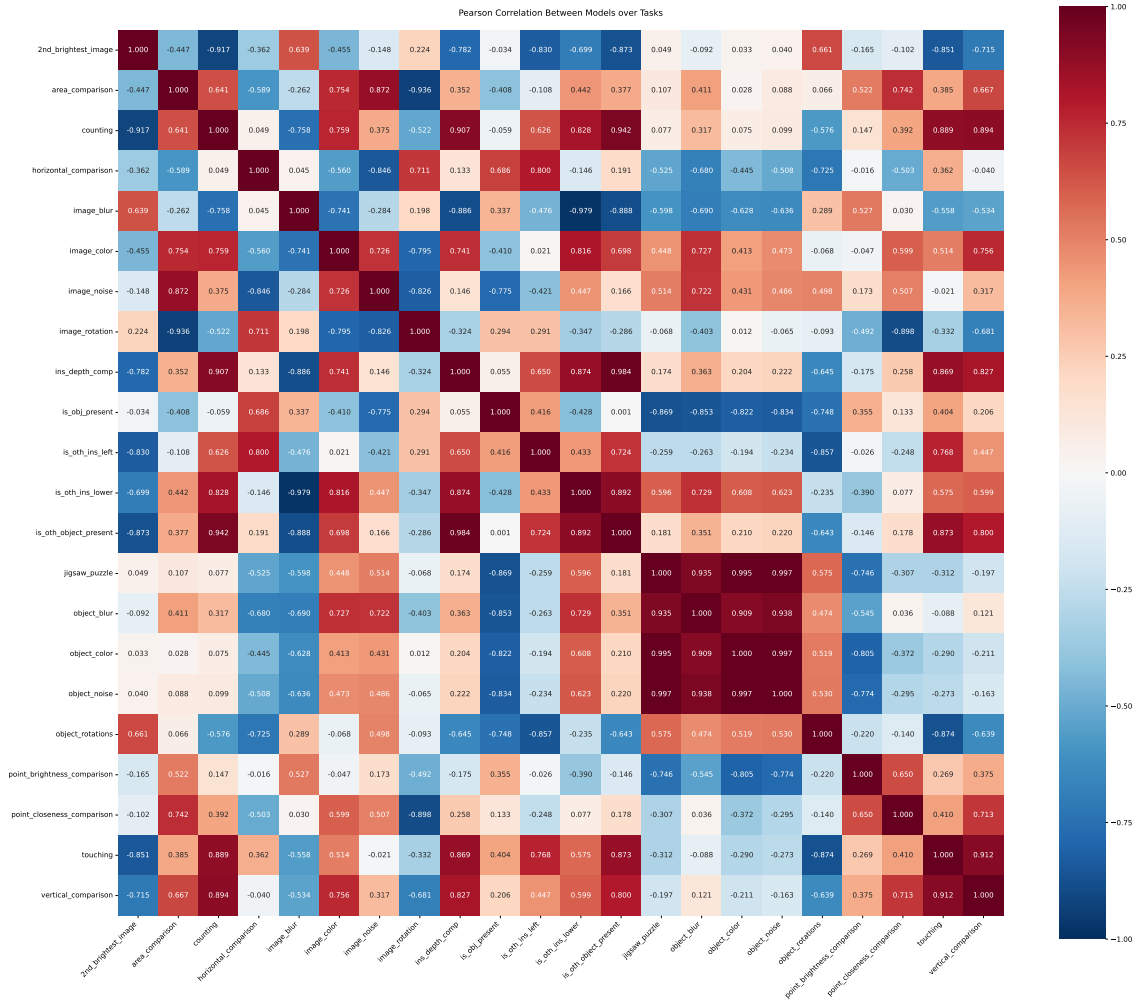


Figure A.27: **Food Dataset: Task-performance correlations highlight shared visual capabilities.** Heatmap of Pearson correlations between task accuracies within the Animals dataset (top-5 models); coefficients range from -1 (anti-correlated) to 1 (highly correlated).

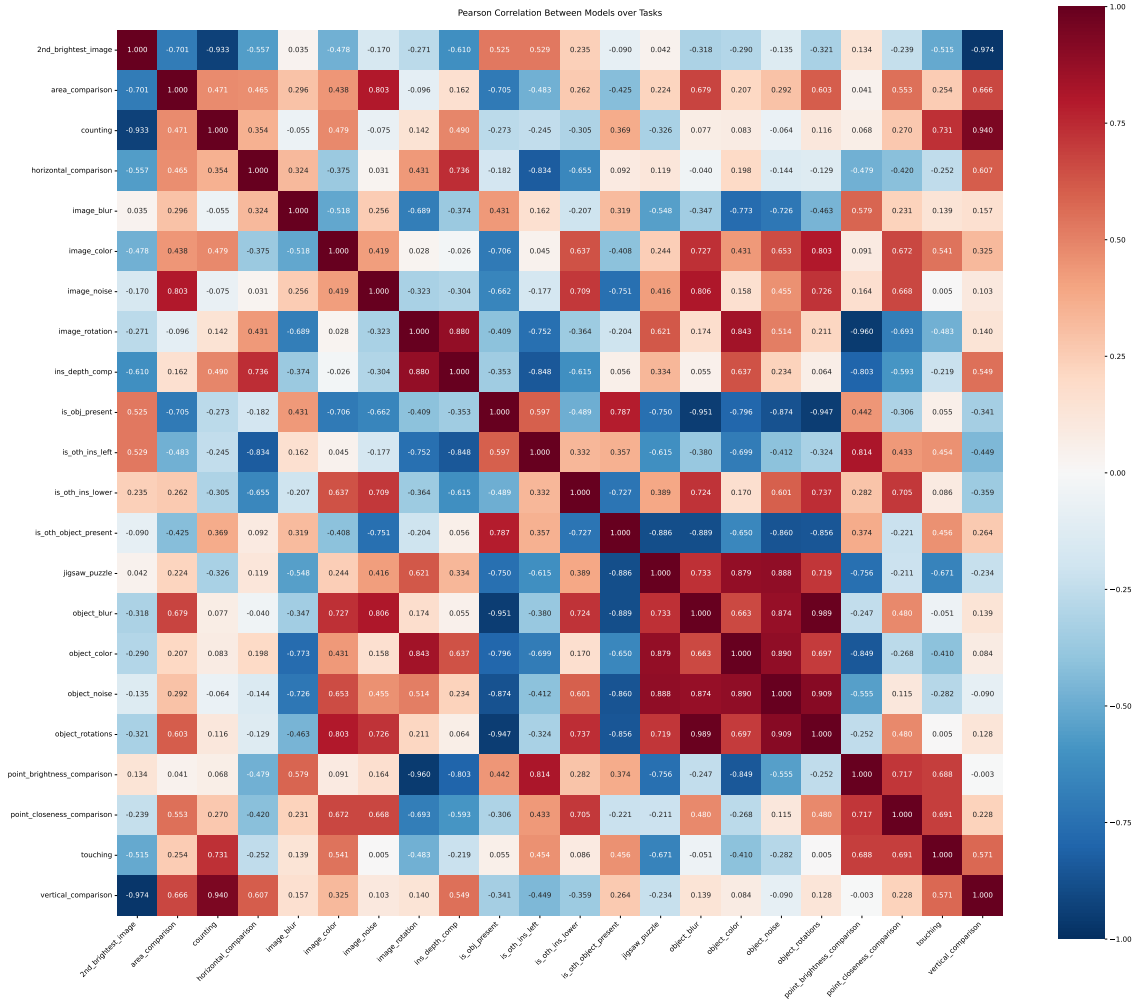


Figure A.28: **Vehicles Dataset: Task-performance correlations highlight shared visual capabilities.** Heatmap of Pearson correlations between task accuracies within the Animals dataset (top-5 models); coefficients range from -1 (anti-correlated) to 1 (highly correlated).

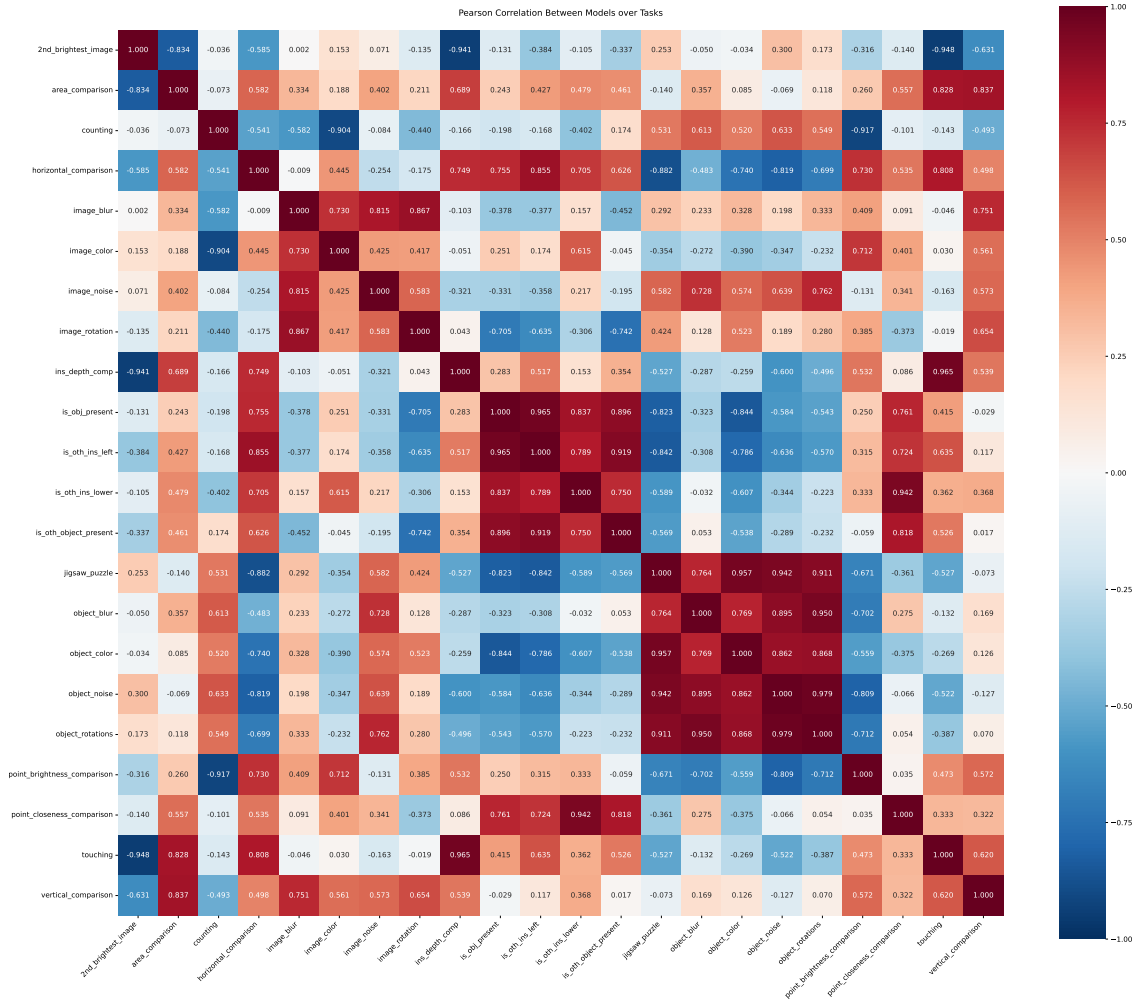


Figure A.29: **Wildlife Dataset: Task-performance correlations highlight shared visual capabilities.** Heatmap of Pearson correlations between task accuracies within the Animals dataset (top-5 models); coefficients range from -1 (anti-correlated) to 1 (highly correlated).

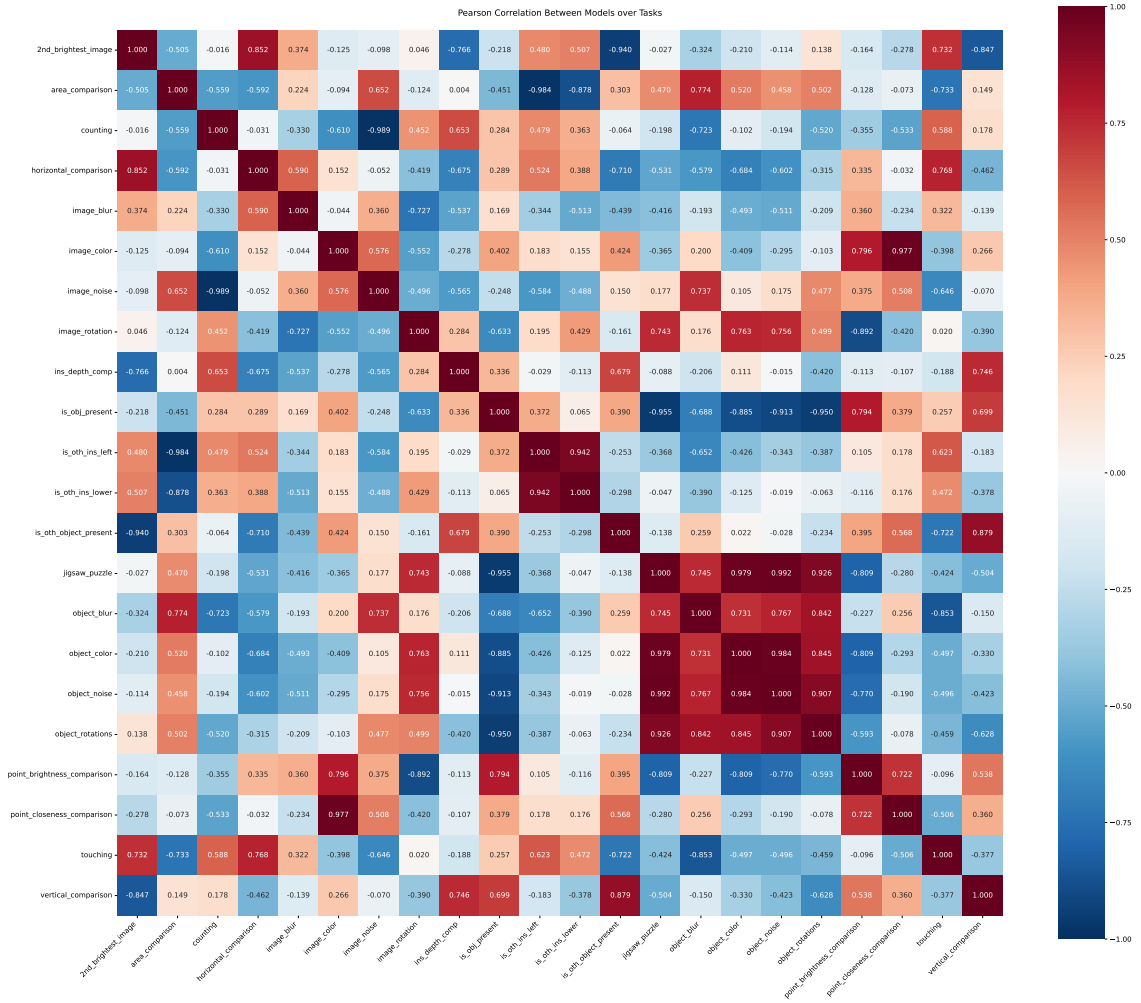


Figure A.30: **Kitchen Dataset: Task-performance correlations highlight shared visual capabilities.** Heatmap of Pearson correlations between task accuracies within the Animals dataset (top-5 models); coefficients range from -1 (anti-correlated) to 1 (highly correlated).

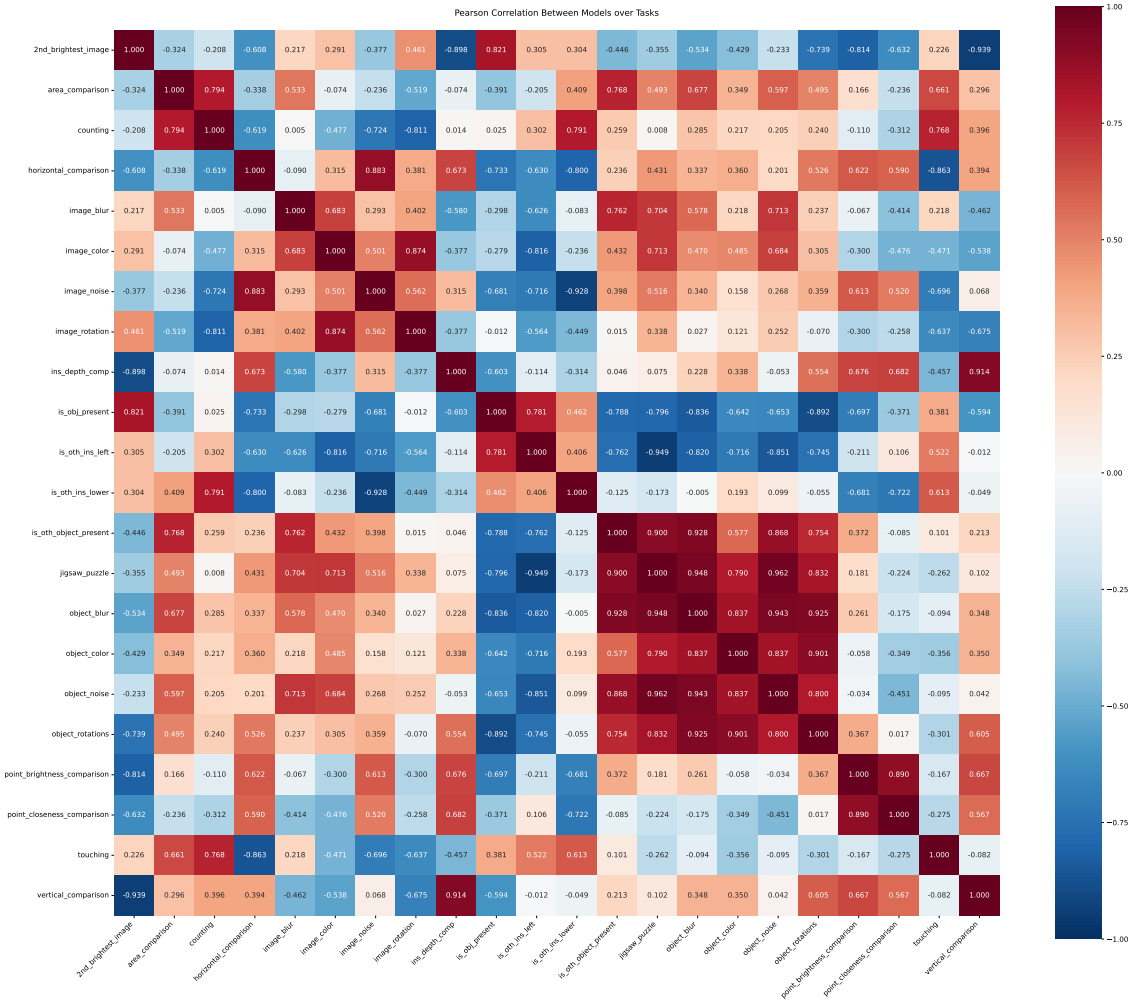


Figure A.31: **Kitti Dataset: Task-performance correlations highlight shared visual capabilities.** Heatmap of Pearson correlations between task accuracies within the Animals dataset (top-5 models); coefficients range from -1 (anti-correlated) to 1 (highly correlated).