BRIDGING VISION LANGUAGE MODEL (VLM) EVALUATION GAPS WITH A FRAMEWORK FOR SCALABLE AND COST-EFFECTIVE BENCHMARK GENERATION

Tim Rädsch[1 2 3 + †] , Leon Mayer[1 +] Simon Pavicic[1] A. Emre Kavur[1 2] Marcel Knopp[1 3]
Barış Öztürk[1], Klaus Maier-Hein[1 2 3] Paul F. Jaeger[1 2] Fabian Isensee[1 2]
Annika Reinke[1 2 †] Lena Maier-Hein[1 2 3 †]
+ core contributor, † project lead [1]German Cancer Research Center (DKFZ) Heidelberg, Germany,
[2]Helmholtz Imaging, DKFZ, Heidelberg, Germany, [3]Heidelberg University, Heidelberg, Germany

ABSTRACT

Reliable evaluation of AI models is critical for scientific progress and practical application. While existing VLM benchmarks provide general insights into model capabilities, their heterogeneous designs and limited focus on a few imaging domains pose significant challenges for both cross-domain performance comparison and targeted domain-specific evaluation. To address this, we propose three key contributions: (1) a framework for the resource-efficient creation of domain-specific VLM benchmarks enabled by task augmentation for creating multiple diverse tasks from a single existing task, (2) the release of new VLM benchmarks for seven domains, created according to the same homogeneous protocol and including 162,946 thoroughly human-validated answers, and (3) an extensive benchmarking of 22 state-of-the-art VLMs on a total of 37,171 tasks, revealing performance variances across domains and tasks, thereby supporting the need for tailored VLM benchmarks. Adoption of our methodology will pave the way for the resource-efficient domain-specific selection of models and guide future research efforts toward addressing core open questions.

1 INTRODUCTION

The reliable and objective performance assessment, i.e., validation of AI models is crucial for both the measurement of scientific progress and translation into practice. Benchmarking for traditional narrow, task-specific AI already comes with numerous challenges (Myllyaho et al., 2021), but validation has proven to be even more complex and error-prone in the emerging field of generalist multimodal foundation models (Schaeffer et al., 2024). In the context of Vision-Language Models (VLMs), one issue that has received limited attention is the heterogeneous and often non-targeted nature of model validation (Tong et al., 2024a;b). Widely used VLM benchmarks span diverse domains and encompass a variety of tasks, providing a broad view of model capabilities across different contexts (Fu et al., 2024b; Liu et al., 2024; Ying et al., 2024; Al-Tahan et al., 2024; Yue et al., 2024).

We identify three key trends that highlight the critical need for personalized benchmarking approaches:

Domain-specific benchmark demand: 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 Appendix A.1. These benchmarks cover a wide range of domains, from autonomous driving to wildlife monitoring, underscoring the need for *domain-specific* benchmarks.

Popular arena platforms do not scale from an individual user's perspective: Arena-style platforms such as Chatbot Arena or WildVision Arena[1] allow users to submit single tasks and rate the outputs of different (anonymized) models. The aggregated user ratings, in turn, can be used for the objective

[1]Chatbot Arena: lmarena.ai/?leaderboard; WildVision Arena: huggingface.co/spaces/WildVision/vision-arena

and comparative assessment of models. While this allows for personalized and domain-relevant evaluation, large-scale assessment from a single user perspective would be cumbersome due to the required annotation effort.

Homogeneous evaluation: Most existing VLM benchmarks (Fu et al., 2024b; Yue et al., 2024; Wang et al., 2024; Zhang et al., 2024c), generally evaluate models using a single question per image. While this can suffice when large datasets are available—allowing for a broad range of tasks—domain experts with smaller, curated datasets face a more significant limitation. From a resource standpoint, image acquisition may also be expensive, and few tasks emerge if there is only one question per image. Furthermore, such an approach provides little insight into whether a VLM truly comprehends broad aspects of an image's semantic content.

Taking these three trends together we conclude that there is a lack of guidance on how to set up a framework that enables personalized, domain-specific benchmarking in a resource-efficient manner. Such a framework must address the scarcity of labeled data, leverage task diversity by systematically generating multiple questions per image, and maintain resource efficiency to ensure accessibility for researchers working in specialized fields, such as wildlife monitoring, or autonomous driving.

In this work, we propose a resource-efficient framework for creating domain-specific VLM benchmarks via task augmentation. Our approach transforms a single type of annotation—instance segmentation—into a diverse set of tasks that test a broad range of perception abilities, such as object counting, occlusion detection, brightness comparison, and more. Specifically, we focus on 2D natural images that either (1) already include instance segmentations or (2) can be annotated using recent advances in semi-automatic labeling tools (e.g., SAM (Ravi et al., 2024)). This approach allows even domains with limited labeled data to efficiently generate custom evaluation tasks. Our main contribution, summarized in Figure 1, is a **resource-efficient framework** for creating domain-specific VLM benchmarks via task augmentation, transforming a single type of annotation (instance segmentation) into a diverse set of tasks. We **apply** this framework **to create seven new domain-specific VLM benchmarks** and comprehensively **evaluate 22 open and closed VLMs on over 37,000 tasks** (for the full model list see Appendix C.1). To establish strong reference points for model evaluation, we collected an additional 162,946 human baseline answers corresponding to 37,171 questions across 1,704 images.

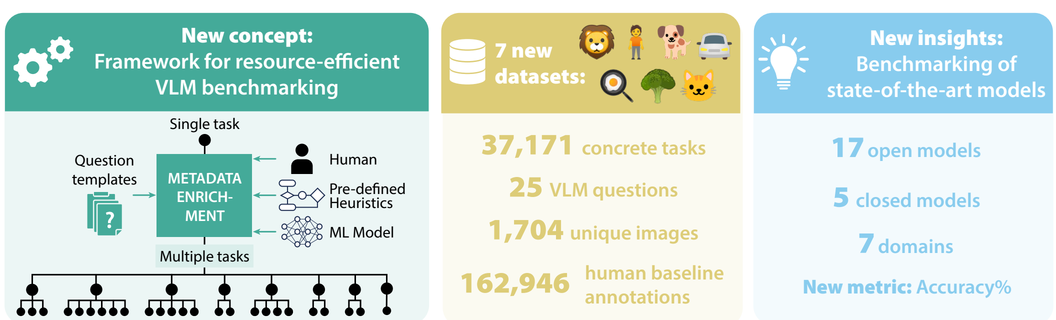
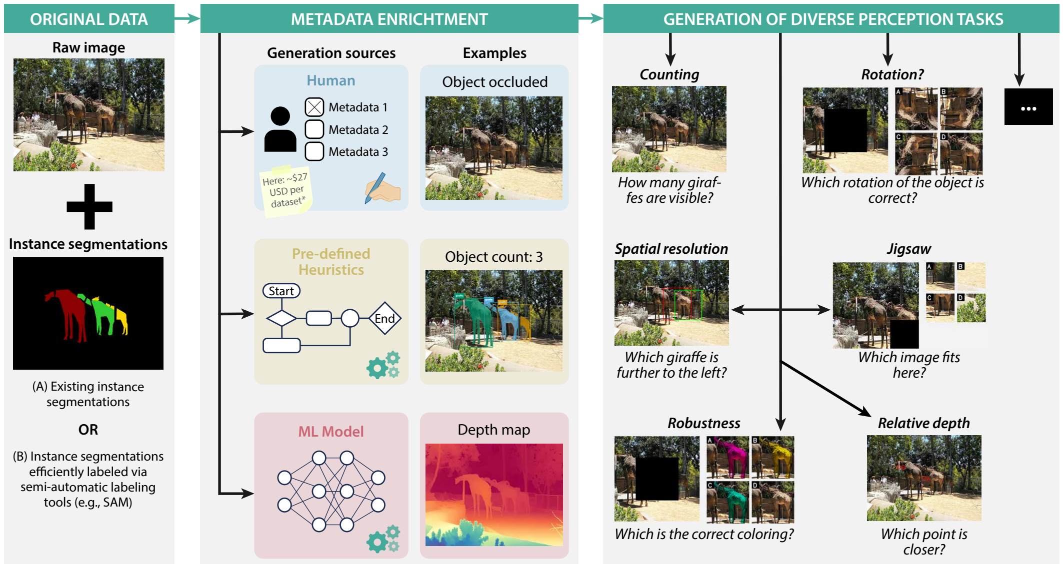


Figure 1: **Summary of contributions**. (1) New concept: We propose a new framework for the resource-efficient creation of domain-specific VLM benchmarks. It is based on the concept of task augmentation designed for creating multiple tasks from a single existing task using metadata annotations from multiple sources (humans, pre-defined heuristics, models). (2) 7 new datasets: We apply our framework 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.



* Annotations for the task generation in our seven datasets cost just $27 USD on average.

Figure 2: **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 from both automatic sources (heuristics and models) and a small number of manual sources (human annotations). This process transforms the initial task into a collection of tasks, enabling resource-efficient and easy to use in-domain benchmarking of general VLM capabilities while maintaining cross-domain comparability.

2 Related Work

2.1 Vision-Language Benchmarks

Recent studies propose a range of evaluation benchmarks for VLMs, varying in size, number, and type of VL capabilities. Examples include Blink (Fu et al., 2024b) and MMBench (Liu et al., 2024) (>3,000 multiple-choice questions each), and MME (Fu et al., 2024a) (Yes/No questions on perception and cognition). The largest benchmarks include MMT-Bench (Ying et al., 2024) (>31,000 questions), MME-RealWorld (Zhang et al., 2024c) (>29,000 image-question pairs), and MMMU (Yue et al., 2024) (>11,500 questions). While these benchmarks cover multiple VL capabilities and domains, they require extensive labeling efforts. For example, MME-RealWorld involved 25 annotators and seven VLM experts, MMMU relied on 50 college students, while MMT-Bench lacks details on annotator numbers. Other benchmarks focus on much smaller question sets (Chen et al., 2024; Yu et al., 2024), integrating multiple existing benchmarks (Jiang et al., 2024; Al-Tahan et al., 2024), or collecting individual human preferences (Lu et al., 2024; Xu et al., 2023). Tong et al. (2024a) present a critical examination of multimodal LLM benchmarks.

Despite the variety of datasets and tasks, a resource-efficient and generalizable approach that enables extensive evaluation of VLMs across multiple (domain-specific) tasks is still lacking. Our framework addresses this gap by empowering users to create domain-specific VLM perception benchmarks from just a few images.

2.2 Task Augmentation and Metadata

Task augmentation refers to generating multiple diverse tasks from a single existing task (Muennighoff et al., 2023). While task augmentation has been addressed from various directions (Johnson et al., 2017; Zhang et al., 2024a; Zamir et al., 2018a; Wang et al., 2023; 2024; Kuznetsova et al., 2020; Krishna et al., 2017), an easy to use framework for evaluating VLMs by domain users on their own images is still missing. The closest works to ours are Zhang et al. (2024a) and Zhang et al. (2024b), which programmatically generate benchmarks using a library of visual assets and task templates. A

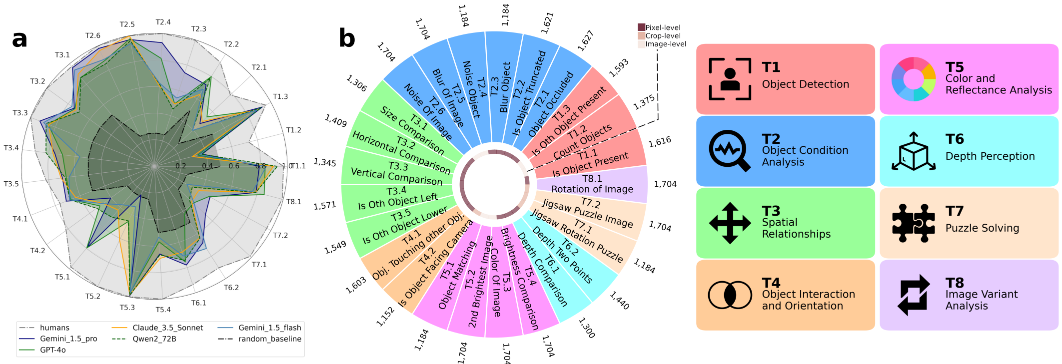


Figure 3: **Our framework yields a diverse set of tasks.** (a) The spider diagram illustrates high Accuracy variability across tasks for the VLMs. We present the results of all the best ranked models while a comprehensive performance summary for all 22 tested models can be found in Appendix C.7. (b) Based on a single image with instance segmentations, our framework enables the generation of 25 tasks from eight different vision-language categories, ranging from pixel-level to image-level perception.

comprehensive comparison to other task augmentations works and their applicability is provided in Appendix A.5.

2.3 Resource-efficient VLM benchmarking

Most existing benchmarks often focus on performance metrics without considering the human and computational resources required to generate a benchmark (see, e.g., (Fu et al., 2024b; Liu et al., 2024)). The work that has been done on efficient benchmarking has been focused in the realm of unimodal language models (Polo et al., 2024; Perlitz et al., 2023). An exception has been Ging et al. (2024), who investigated the automatic creation of VLM benchmarks from classification datasets. Nevertheless, the increasing prominence of VLMs in research and industry (Li et al., 2024; Yang et al., 2023) is not yet reflected in efforts to increase efficiency during benchmark creation.

3 Methods

3.1 Framework for resource-efficient in-domain benchmarking

The framework for resource-efficient in-domain benchmarking is depicted in Figure 2. Starting with domain images that include instance segmentations (existing or created with semi-automatic labeling tools, such as SAM (Ravi et al., 2024)), metadata for each image is acquired from multiple sources (humans, pre-defined heuristics, and models) to transform the single task into a collection of perception tasks.

For our seven new datasets, we use existing instance segmentation as the core perceptual task to generate the diverse set of VLM benchmark tasks depicted in Appendix A.4 (examples in Figure 5 and more detailed in Appendix B.2).

The metadata enrichment is derived from three sources:

1) Human annotators were used to generate information that cannot be extracted from the existing annotations or using established models. To this end, we outsourced annotations to a professional annotation company (Quality Match GmbH in Heidelberg). Specifically, human raters were tasked with determining the presence of occlusion and truncation in the images. Furthermore, they were asked to assess the direction in which the objects were facing. These annotations cost ∼27 USD on average with a total turnover time of two days.

2) 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.

Domain	Icon	#Images	#Objects	#Tasks	#Human 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**	**37,171**	**162,946**

Table 1: **Dataset statistics across different domains.** The table presents the total number of images, objects, tasks, and human annotations across all domains.

3) An existing depth foundation model, Depth Anything v2 (Yang et al., 2024), was used to generate depth maps for each image.

3.2 Seven new datasets from diverse domains

We applied our proposed framework to images from seven different domains. Overall, the input images and instance segmentations for our framework were extracted from KITTI (Geiger et al., 2012), COCO (Lin et al., 2014), and COCONut (Deng et al., 2024). In summary, we added 300,000 metadata annotations to a total of 1,704 images across seven domains. This includes 15 annotations per object (e.g. occlusion, relative_size, segmask_touches_segmask, or average_depth). For truncation, occlusion, and direction, we obtained up to five annotations per object from human annotators (UI example is displayed in Appendix A.2). Early stopping was applied when four annotators reached a consensus. The complete list is provided in Appendix Table 4.

The metadata were then used to define a set of 25 different VLM tasks (see Figure 3), including six tasks concerning the entire image, 13 related to individual objects, and six focused on object pairs.

Setup for automatic task processing after metadata extraction: 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. 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. Overall, our objective was to generate as many of the 25 different tasks as possible for each image. No LLMs or VLMs were used for task generation, as these methods are prone to injecting hallucinations (Wang et al., 2023; 2024). We prioritized quality and reliability instead.

Human ambiguity baseline: To rate the difficulty and ambiguity for each of the 37,171 tasks, we further acquired annotations from six human raters per image. We implemented early stopping if four raters reached agreement on a task. Overall, this resulted in 162,946 human reference annotations. An overview of the resulting datasets is provided in Table 1and exemplary images for all generated datasets are included in Appendix B.

3.3 Benchmarking strategy

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. For example, Accuracy is a prevalence-dependent metric, meaning that results should not be compared across datasets. To address this bottleneck, we fully homogenized our benchmarking framework using the proposed framework.

	Overall	Wildlife 🦁	Animals 🐕	Kitti 🐱	Person 🧍	Vehicles 🚘	Food 🥦	Kitchen 🍳
Human	93.7	93.4	93.9	94.6	95.2	92.4	93.6	92.6
Gemini_1.5_pro	72.4	74.6	75.4	78.0	70.8	71.7	70.0	66.6
GPT-4o	69.8	71.4	71.4	76.2	69.0	69.3	67.0	64.4
Claude_3.5_Sonnet	69.0	73.7	74.3	72.4	65.3	67.8	65.7	63.8
Qwen2_72B	68.8	70.8	75.0	74.6	61.7	68.2	66.2	64.7
Llama_3.2_90B	65.9	71.3	70.2	68.6	64.6	63.1	62.9	60.8
Gemini_1.5_flash	65.7	71.5	70.2	70.8	60.2	65.2	61.1	61.0

Table 2: **The rankings of models differ strongly across the tested domains.** Model Accuracies across different generated datasets. 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 Appendix Table 9. Note that Accuracy does not account for shared images between questions; this issue is addressed in Figure 4.

Model selection: We selected 22 frontier and open VLMs of various sizes and from various providers and sources, as illustrated in Appendix C.1. The oldest model was released in January 2024, while the most recent one included was released at the end of September 2024.

Benchmarking workflow: To ensure fair and consistent evaluation of all selected VLMs, we developed a standardized benchmarking workflow applied uniformly across all models. We assessed them in a zero-shot setting without any additional fine-tuning or domain-specific training. 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 human evaluators in our department. Through four rounds of refinement, we adjusted the prompts until all four human evaluators consistently agreed on their interpretation. 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.

VLM tasks: We evaluated the models on a comprehensive set of 25 tasks derived from our task augmentation framework (overview in Figure 3, full list in Appendix A.4 and examples per dataset in Appendix B). Each task was associated with specific evaluation criteria and standardized prompts. 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. This attention to detail ensured that the evaluation was both rigorous and reproducible.

Metrics and rankings: Choosing an adequate strategy for performance assessment is far from trivial and a research topic of its own (Maier-Hein et al., 2024; Reinke et al., 2024). In this work, we were specifically interested in relative performance differences rather than in the specific ability of VLMs to serve a specific task. 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:

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

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

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

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 Experiments and Results

The primary purpose of our experiments was to showcase the benefit of our task augmentation approach (sec. 4.1). 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 ambiguity (sec. 4.2). Finally, we leveraged our concept and data to explore the capabilities of the most recent open and closed VLMs (sec. 4.3.).

4.1 Benefit of the proposed framework

Figure 2 shows aggregated performance values for all models, separated by imaging domain. As the tasks and prompts were homogenized, the results clearly indicate that performance 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.

Furthermore, as shown in Figure 3a, 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”, (4), T2.1 “Object Occlusion Detection”, and (5) T5.2 “Second Brightest Image Selection”. The easiest task on average was T1.3 “Additional Object Presence Detection” (see Figure 24).

4.2 Human Ambiguity

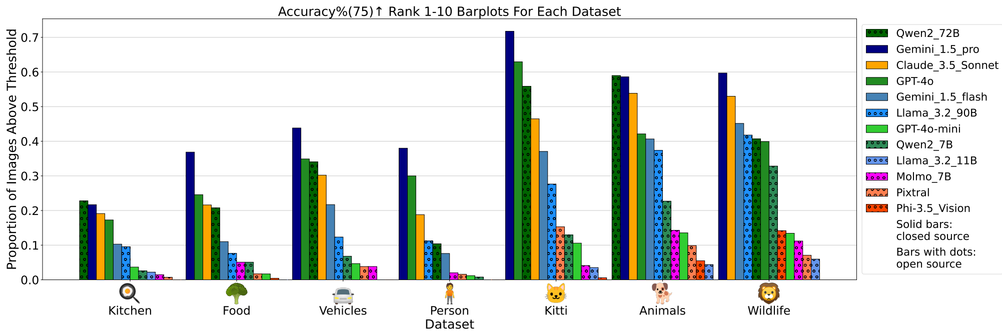
As demonstrated in Appendix C.3, there is a high discrepancy in task rankings between humans and models. While the ”Jigsaw Puzzle Completion” tasks ranked amongst the most challenging for the models, humans found ”Object Occlusion Detection” and “Object Touching Detection” to be the most difficult.

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 Appendix subsection C.6. It can be seen that many hard tasks, including the top four, can already be extracted from instance segmentations alone.

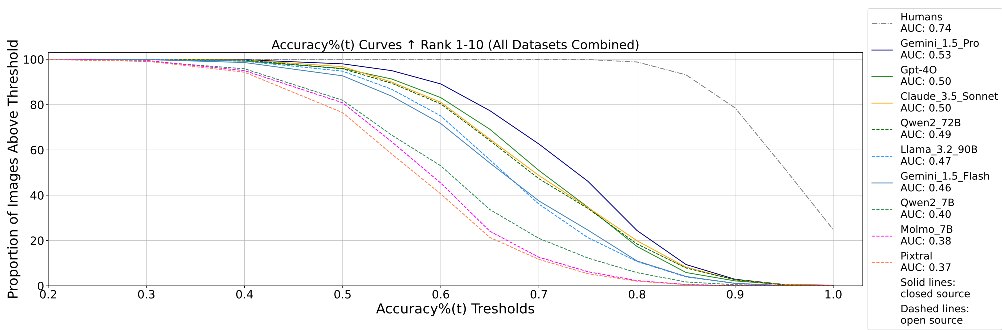
4.3 Insights on current models

Figure 4 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 confirms well-known trends: closed models still outperform across tasks, though open models have notably narrowed the gap. In particular, Qwen2 72B stands out as the strongest performer among open models. 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, occlusion, and direction-related tasks—counting being particularly problematic. Regarding model scaling, larger variants



(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 1.5 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 C.2.



(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 state-of-the-art model, Gemini_1.5_pro, achieves only 0.53, highlighting significant room for improvement. Only the top 10 models are shown. Curves for all 22 models and datasets are displayed in Appendix C.8.

Figure 4: **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 0.37 to 0.53 for the top 10 models tested.

typically show better performance, with some notable exceptions such as Molmo 7B outperforming Pixtral 12B.

Interesting new findings: The need for specific in-domain evaluation is highlighted by the high performance variability across imaging domains for the same tasks, see Table 2 and Figure 2. The overall best model, Gemini 1.5 Pro, varies between domains from 22% (Kitchen) to 72% (Kitti). Qwen2 72B slightly surpasses Gemini 1.5 Pro on the kitchen and animals datasets but ranks only fifth on the person dataset. Additional insights emerge from model comparisons, with Qwen2 7B consistently outperforming Molmo 7B across most datasets, and Gemini Flash 1.5 showing superior Point Depth Comparison capabilities over Gemini Pro. These results indicate that our newly introduced metric, Accuracy%(t), can effectively capture model performance in a single value.

5 DISCUSSION

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

1) 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

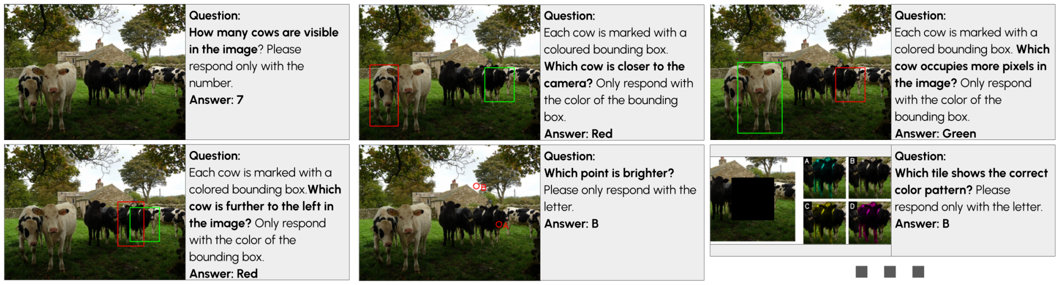


Figure 5: **Our framework yields a diverse set of tasks.** Exemplary tasks that were generated with the framework for a given image. A broad range of examples and errors for each generated dataset is provided in Appendix B.

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 A.6.

2) Seven new openly available datasets: Our seven new datasets will help assess generalist capabilities of future VLMs. Furthermore, we release the six human annotations per task (totaling 162,946 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. Core strengths of our contribution include the broad applicability of our concept, the open dataset and benchmark contribution, and the wide range of state-of-the-art closed and open models investigated.

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 extendable 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 (Reinke et al., 2024). 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.

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. We provide further statistical analysis, such as ranking variability in Appendix C. Model performance showed small variations with prompt phrasing, which we mitigated through iterative testing for consistency. Additionally, our human annotations were performed by professional annotators, which may introduce ambiguity since annotators aim to complete tasks quickly.

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.

Code / Datasets / Human Annotations

Code, datasets, and annotations will be made available.

References

Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. *arXiv preprint arXiv:2408.04810*, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. Coconut: Modernizing coco segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21863–21873, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2024a.

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. *arXiv preprint arXiv:2404.12390*, 2024b.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.

Simon Ging, María A Bravo, and Thomas Brox. Open-ended vqa benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. *arXiv preprint arXiv:2402.07270*, 2024.

Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2, 2024.

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 *CVPR*, 2017.

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 Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis*, 123:32–73, 2017.

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. *Int J Comput Vis*, 128:1956–1981, 2020.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2024.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024.

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. *arXiv preprint arXiv:2308.07124*, 2023.

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, 2021. ISSN 0164-1212. doi: https://doi.org/10.1016/j.jss.2021.111050. URL https://www.sciencedirect.com/science/article/pii/S0164121221001473.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). *arXiv preprint arXiv:2308.11696*, 2023.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

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, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Annika Reinke, Minu D Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A Emre Kavur, Tim Rädsch, Carole H Sudre, Laura Acion, Michela Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature methods*, 21(2):182–194, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024a. URL https://arxiv.org/abs/2406.16860.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

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. *arXiv preprint arXiv:2311.07574*, 2023.

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. *arXiv preprint arXiv:2409.12953*, 2024.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024.

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 CVPR*, 2024.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018b.

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, 2024a. URL `https://arxiv.org/abs/2406.11775`.

Jieyu Zhang, Le Xue, Linxin Song, Jun Wang, Weikai Huang, Manli Shu, An Yan, Zixian Ma, Juan Carlos Niebles, Caiming Xiong, et al. Provision: Programmatically scaling vision-centric instruction data for multimodal language models. *arXiv preprint arXiv:2412.07012*, 2024b.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024c.

We provide further detail on i) the ressource-efficient framework, ii) the seven generated domain-specific datasets, and iii) the benchmarking insights and model evaluations.

Impact Statement

This paper advances Machine Learning by enabling researchers to benchmark with their own data on a minimal budget. All human annotations were sourced from a reputable company following ethical guidelines.

Contents

A RESOURCE-EFFICIENT VLM BENCHMARKING FRAMEWORK

A.1 CVPR 2024 PAPER ANALYSIS

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 3: **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.

We analyzed all papers from CVPR 2024 using three different large language models (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.

A.2 Example Of Human Generated Metadata

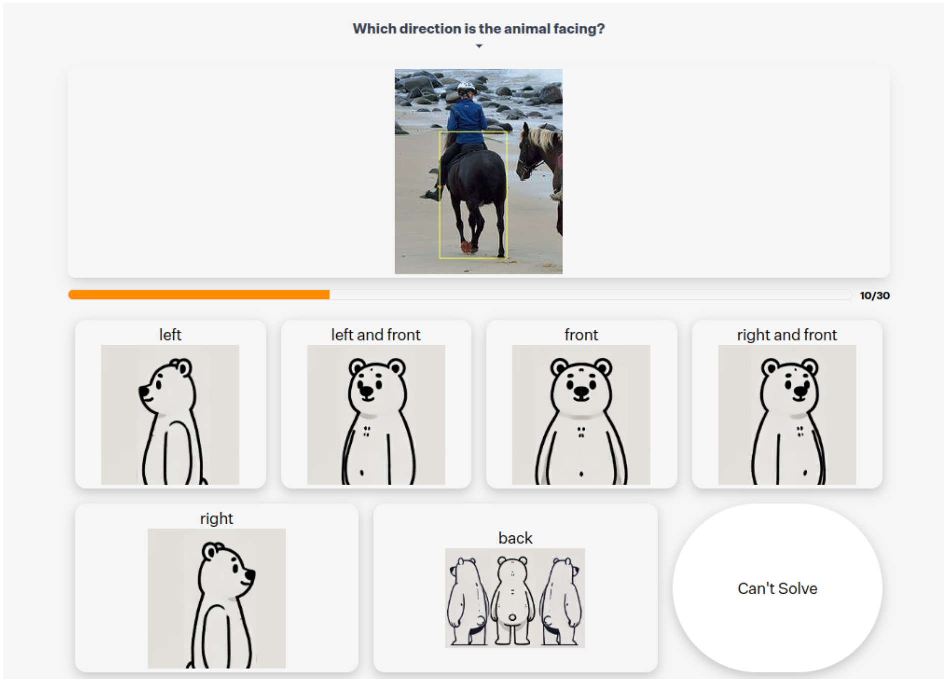


Figure 6: **Example of human-generated metadata enriching object annotations in initialization tasks.** These annotations demonstrate the process of enriching objects with human-provided metadata during the initial setup phase. The visual displays in the instructions enable consistent annotations (Rädsch et al., 2023).

A.3 Metadata Sources

Human Raters	
Attribute	**Description**
Occluded	Object occluded or fully visible (other object in front)
Truncated	Object truncated or fully visible (edge of image)
Direction	Direction the object is facing
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: **Overview of metadata sources used for enriching instance segmentation datasets.** Metadata was created from existing annotations, specialized models, or manually annotated by human raters.

A.4 VLM Tasks Overview

Here we present the VLM tasks overview and its corresponding meta categories in Figure 7. Further information on each task is provided in Table 5 on the next page.

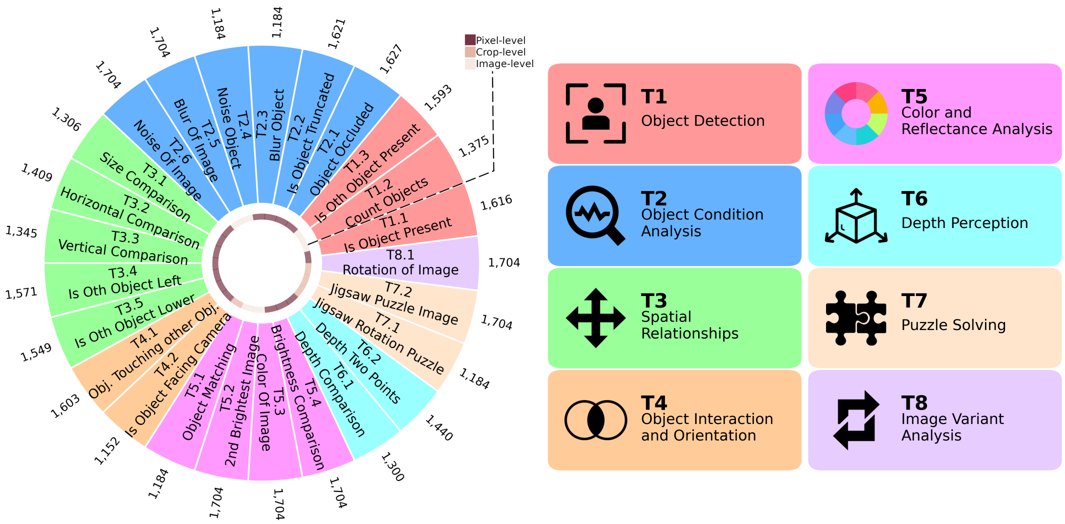


Figure 7: **Our framework yields a diverse set of tasks.** Based on a single image with instance segmentations, our framework enables the generation of 25 tasks from eight different vision-language categories, ranging from pixel-level to image-level perception.

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	Is Object Occluded	Determines if the specified object is partially or fully occluded.	Quiz (A/B/C/D)
T2.2	Is Object Truncated	Determines if the specified object is truncated in the image frame.	Binary
T2.3	Blur Object	Determines whether an object is blurred	Quiz (A/B/C/D)
T2.4	Noise Object	Determines whether an object contains noise	Quiz (A/B/C/D)
T2.5	Blur Of Image	Determines which image variant is least blurred	Quiz (A/B/C/D)
T2.6	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
T4.2	Is Object Facing Camera	Determines if the object is facing the camera	Quiz (A/B/C/D)
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 5: **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 B.1.

A.5 Task Augmentation Methods Comparison

Metric	CLEVR	Task Me Anything	Taskonomy	Wang 2023	JourneyBench	ProVision	Ours
	Johnson et al. (2017)	Zhang et al. (2024a)	Zamir et al. (2018b)	Wang et al. (2023)	Wang et al. (2024)	Zhang et al. (2024b)	**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 6: **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 (Johnson et al., 2017; Zhang et al., 2024a; Zamir et al., 2018b; Wang et al., 2023; 2024) is provided in Table 6.

We deliberately excluded large datasets primarily designed for task generation, such as Visual Genome (Krishna et al., 2017) and the Open Images Dataset (Kuznetsova et al., 2020), 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.

A.6 Computational and Monetary Cost

Domain	Icon	Metadata Enrichment			Perception Task Generation (in seconds)
		Pre-defined Heuristics (in seconds)	ML Model (in seconds)	Human Annotations for task generation	
Wildlife	🦁	15	50	$\sim$ 27$ USD on average per dataset (2-day average turnover)	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 7: **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. Annotation costs averaged $27 per dataset, calculated based on the number of objects and manually added metadata points. The turnover time for task annotations was two days. All computations were performed on an RTX 3090 GPU and Ryzen 9 5900X CPU.

B DOMAIN-SPECIFIC DATASETS

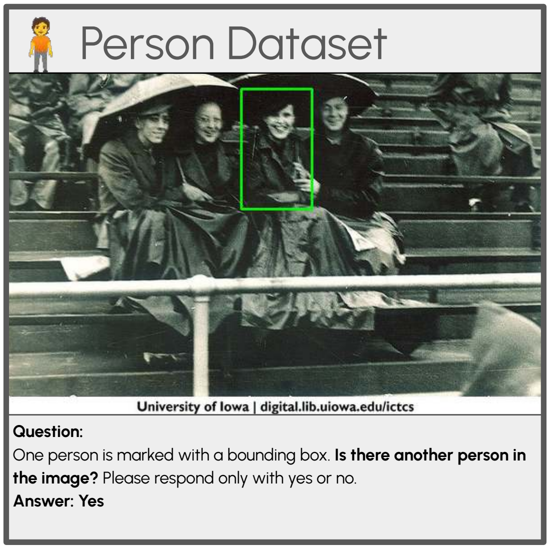
B.1 Animals Dataset Examples

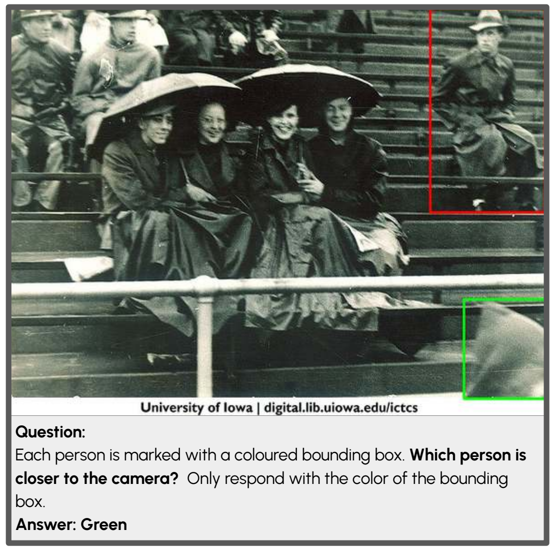
Question:
Which point is brighter? Please only respond with the letter.
Answer: B

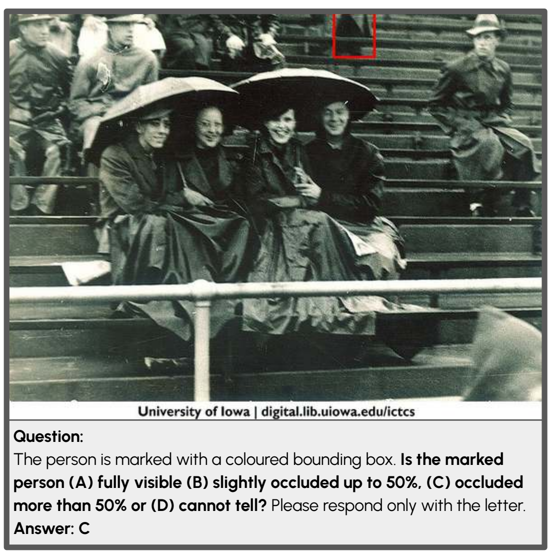
Question:
Which tile fits best in the image? Please respond only with the letter..
Answer: A

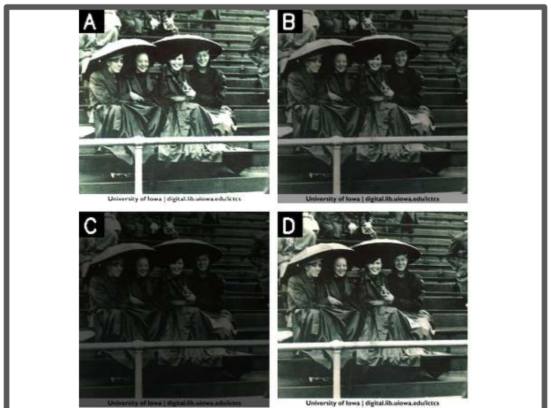
Question:
Which tile shows the correct color pattern? Please respond only with the letter.
Answer: B

B.2 Person Dataset Examples



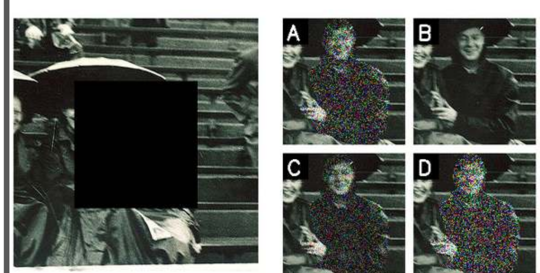






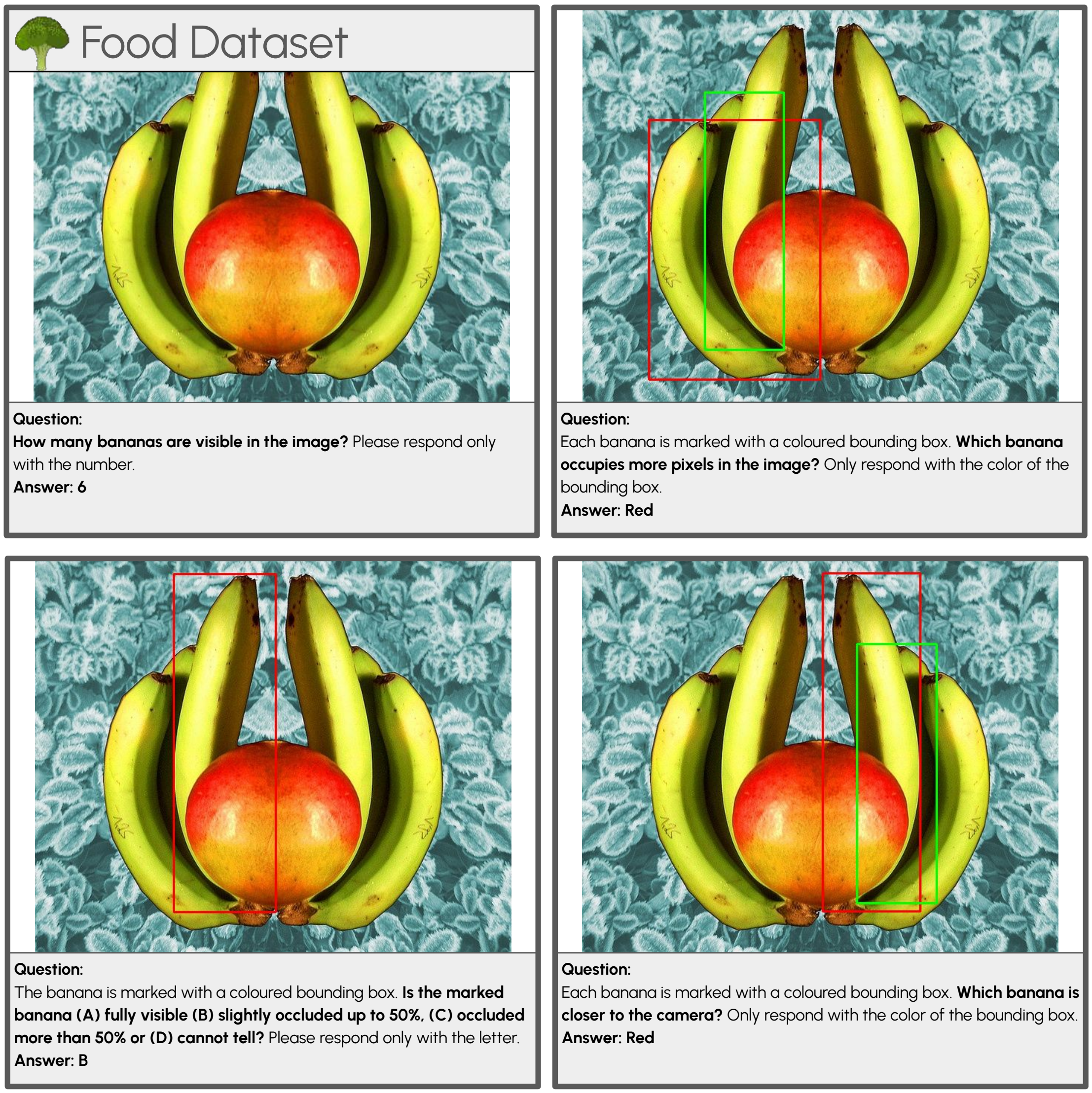
Question:
Which image variant is the second brightest version of the image? Please respond only with the letter.
Answer: D

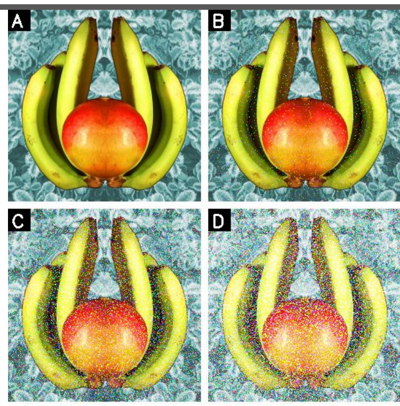
Question:
Which point is brighter? Please respond only with the letter.
Answer: A



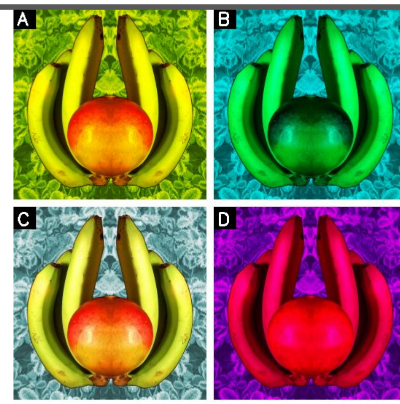
Question:
Which tile fits best in the image? Please respond only with the letter.
Answer: B

B.3 Food Dataset Examples





Question:
Which image variant is not corrupted? Please respond only with the letter.
Answer: A



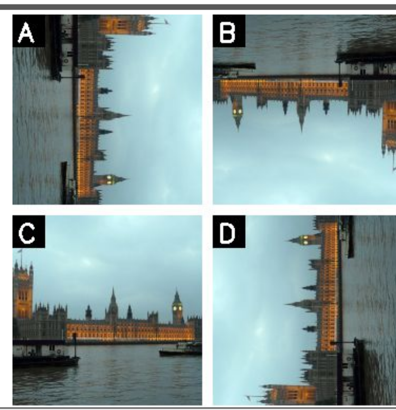
Question:
Which image variant is not corrupted? Please respond only with the letter.
Answer: C

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

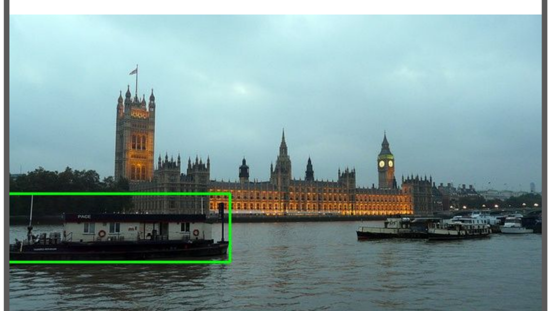
B.4 Vehicles Dataset Examples

Vehicles Dataset

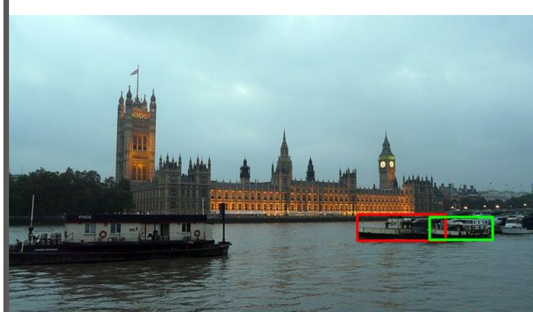
Question:
The boat is marked with a coloured bounding box. **Is the marked boat truncated?** Truncated means that the object is cut off at the edge of the image. Please respond only with yes or no.
Answer: No



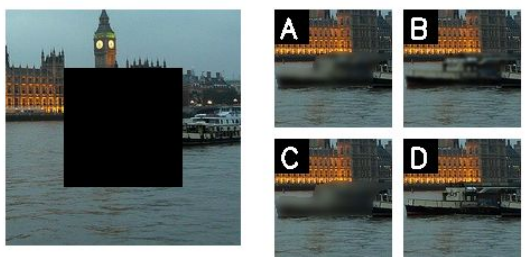
Question:
Which image variant is not rotated? Please respond only with the letter.
Answer: C



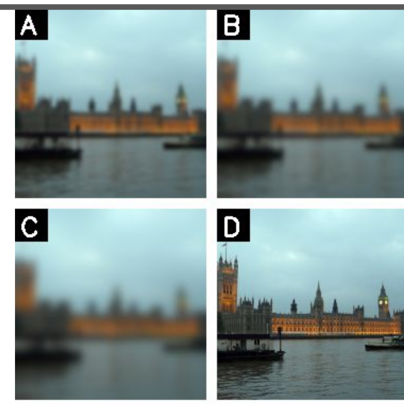
Question:
The boat is marked with a coloured bounding box. I**s the marked boat (A) fully visible (B) slightly occluded up to 50%, (C) occluded more than 50% or (D) cannot tell?** Please respond only with the letter.
Answer: A



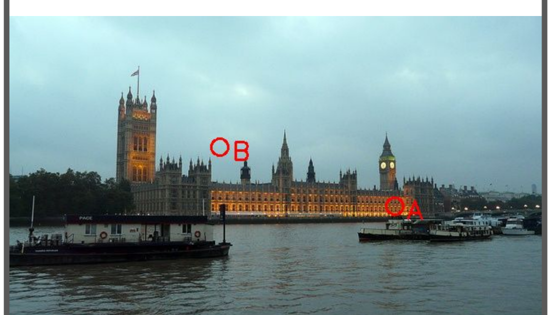
Question:
Each boat is marked with a coloured bounding box. **Which boat is further to the left in the image?** Only respond with the color of the bounding box.
Answer: Red



Question:
Which tile fits best in the image? Please respond only with the letter.
Answer: D



Question:
Which image variant contains the smallest amount of blur? Please respond only with the letter.
Answer: D

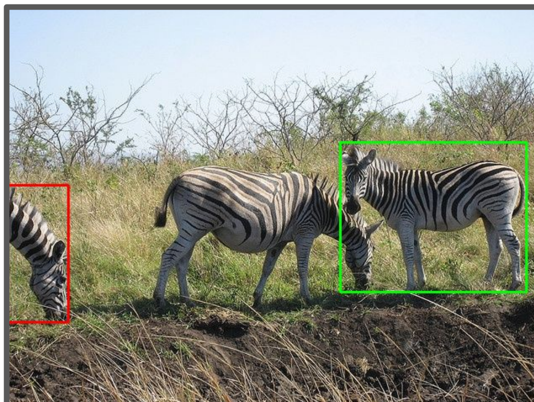


Question:
Which point is closer? Please respond only with the letter.
Answer: A

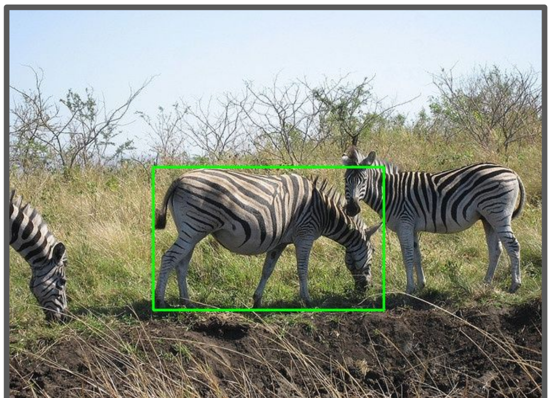
B.5 Wildlife Dataset Examples

Wildlife Dataset

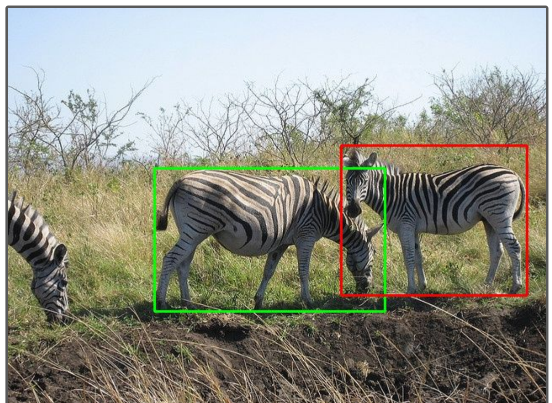
Question:
The coloured bounding box marks an area of interest. **Is there a giraffe in the bounding box?** Please respond only with yes or no.
Answer: No



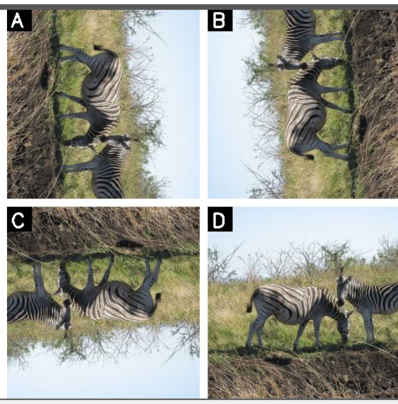
Question:
Each zebra is marked with a coloured bounding box. **Which zebra is closer to the bottom of the image?** Only respond with the color of the bounding box.
Answer: Red



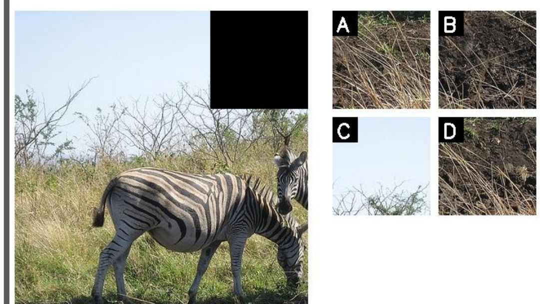
Question:
The zebra is marked with a colored point. **Does the zebra touch another zebra?** Touching means that the pixels of one zebra are directly next to the pixels of another zebra. Please respond only with yes or no.
Answer: Yes



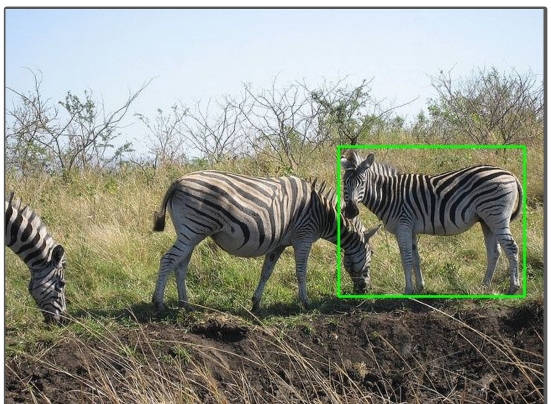
Question:
Each zebra is marked with a coloured bounding box. **Which zebra is closer to the camera?** Only respond with the color of the bounding box.
Answer: Green



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

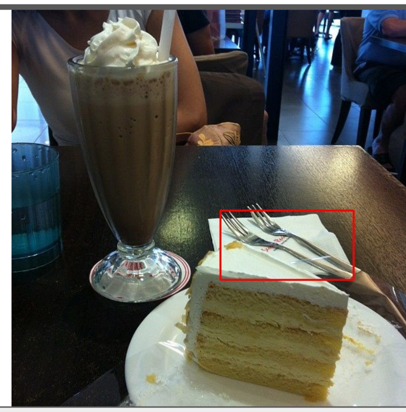


Question:
The zebra is marked with a coloured bounding box. **Is the marked zebra truncated?** Truncated means that the object is cut off at the edge of the image. Please respond only with yes or no.
Answer: No

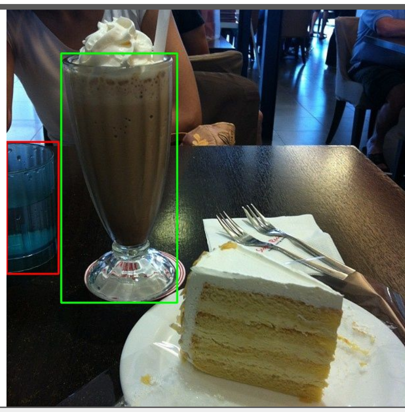
B.6 KITCHEN DATASET EXAMPLES

Kitchen Dataset

Question:
How many forks are visible in the image? Please respond only with the number.
Answer: 2

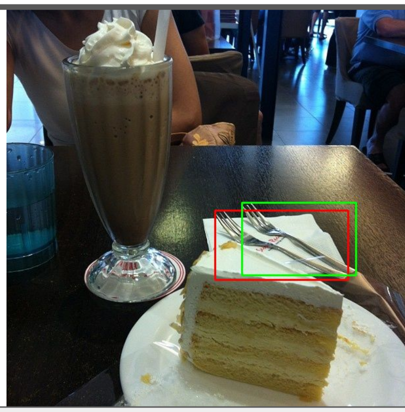


Question:
The fork is marked with a colored point. **Does the fork touch another fork?** Touching means that the pixels of one fork are directly next to the pixels of another fork. Please respond only with yes or no.
Answer: Yes

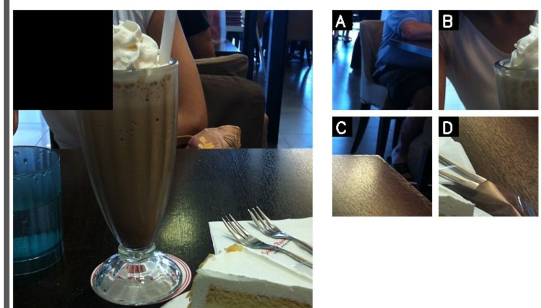


Question:
Each cup is marked with a coloured bounding box. **Which cup is closer to the camera?** Only respond with the color of the bounding box.
Answer: Green

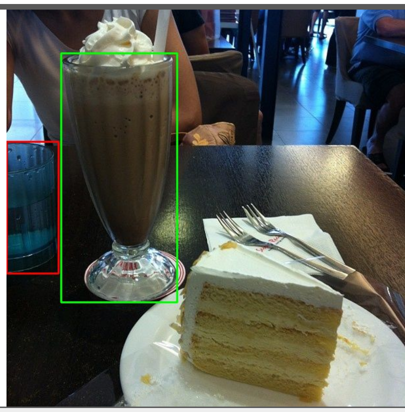
Question:
Which point is closer? Please respond only with the letter.
Answer: B



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



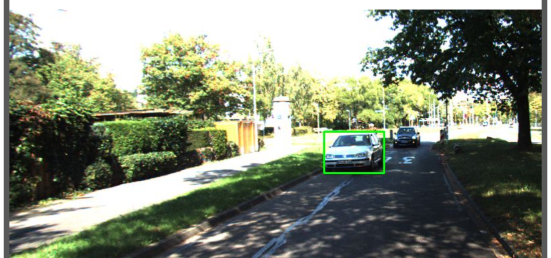
Question:
Which tile fits best in the image? Please respond only with the letter..
Answer: B



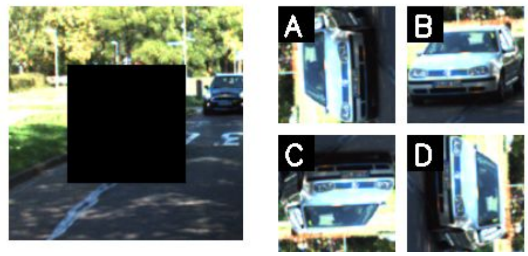
Question:
Each cup is marked with a coloured bounding box. **Which cup occupies more pixels in the image?** Only respond with the color of the bounding box.
Answer: Green

B.7 Kitti Dataset Examples

🐱 Kitti Dataset



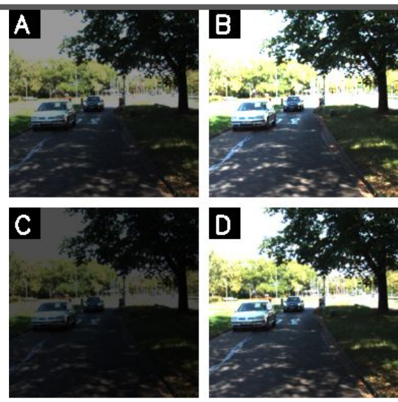
Question:
The coloured bounding box marks an area of interest. **Is there an car in the bounding box?** Please respond only with yes or no.
Answer: Yes



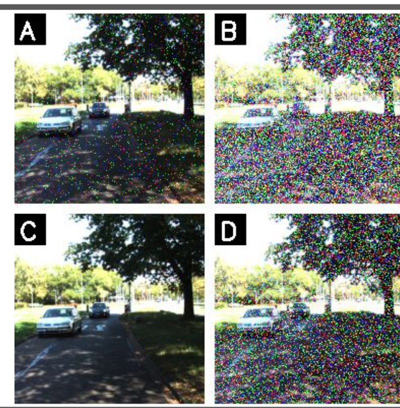
Question:
Which of the four tiles fits best in the image? Please respond only with the letter.
Answer: B

Question:
The car is marked with a coloured bounding box. **From the cameras perspective, is the car (A) facing somehow towards the camera, (B) facing left, (C) facing backwards away from the camera or (D) facing right?** Please respond only with the letter.
Answer: A

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 image variant is not corrupted? Please respond only with the letter.
Answer: C

Question:
Which point is brighter? Please respond only with the letter.
Answer: B

B.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 8 and Figure 9 we display some tasks that triggered the safety settings.

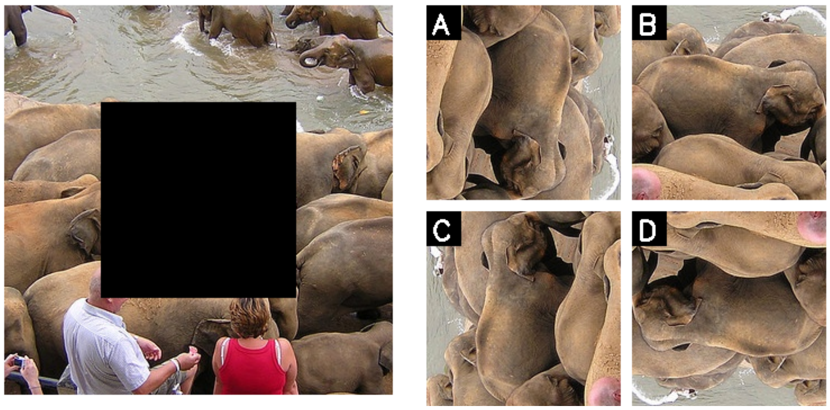


Figure 8: **Safety systems in vision-language models can be triggered by benign inputs.** Example showing an image that activated Google Gemini's content filtering mechanisms despite containing no harmful content.

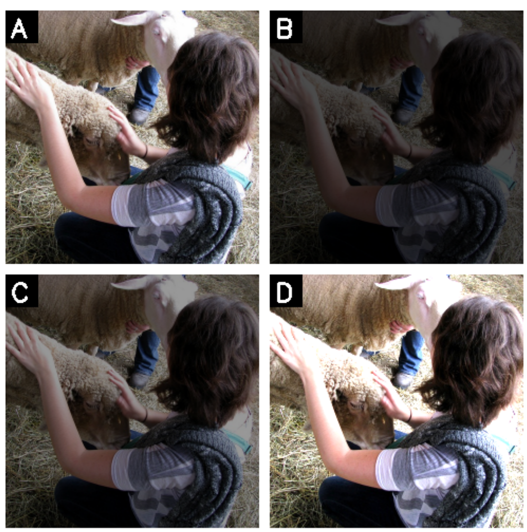


Figure 9: **Safety systems in vision-language models can be triggered by benign inputs.** Example showing an image that activated Google Gemini's content filtering mechanisms despite containing no harmful content.

C BENCHMARKING INSIGHTS AND MODEL EVALUATIONS

C.1 VLM MODEL OVERVIEW

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 8: **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.

C.2 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 10. To increase interpretability of single models across datasets, Figure 11 displays all 22 models as line plots.

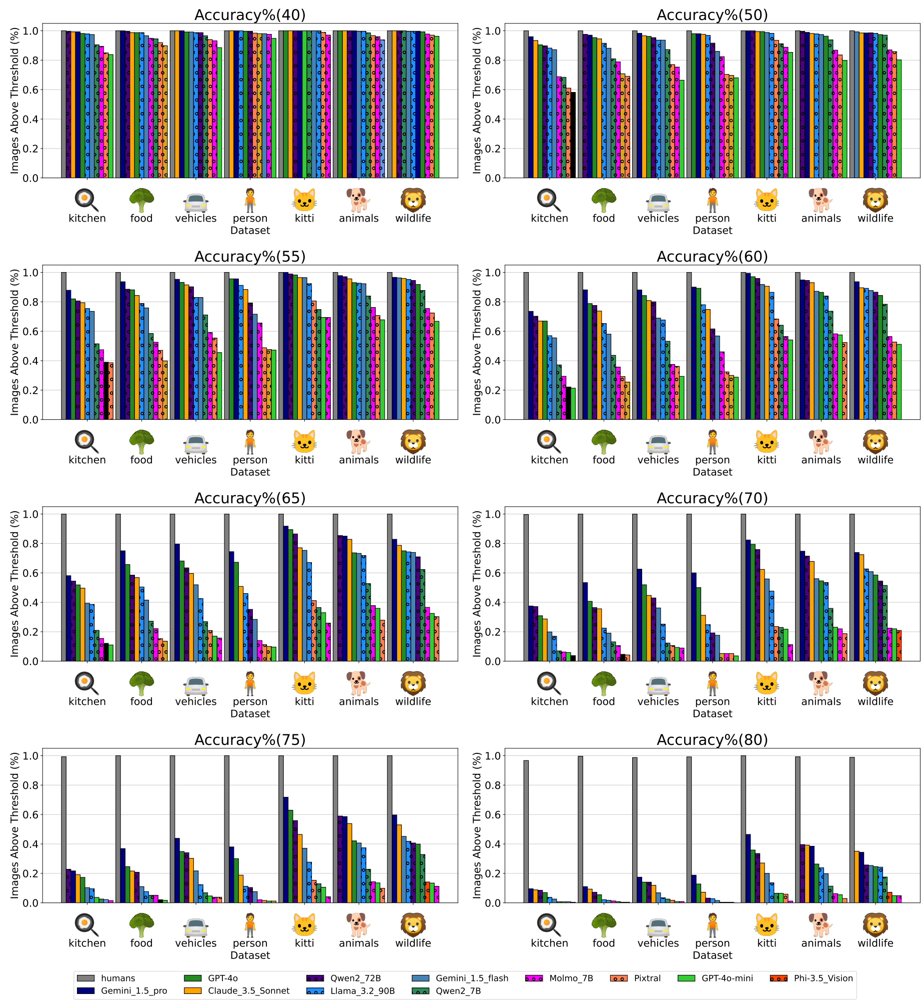


Figure 10: **Accuracy% (t) ↑thresholds for the top 10 models across all datasets**. Humans (solid grey bar) consistently achieve near-perfect scores of 1 across all displayed thresholds and datasets. Performance varies significantly across imaging domains, with dotted lines representing open-source models and solid lines indicating closed-source models.

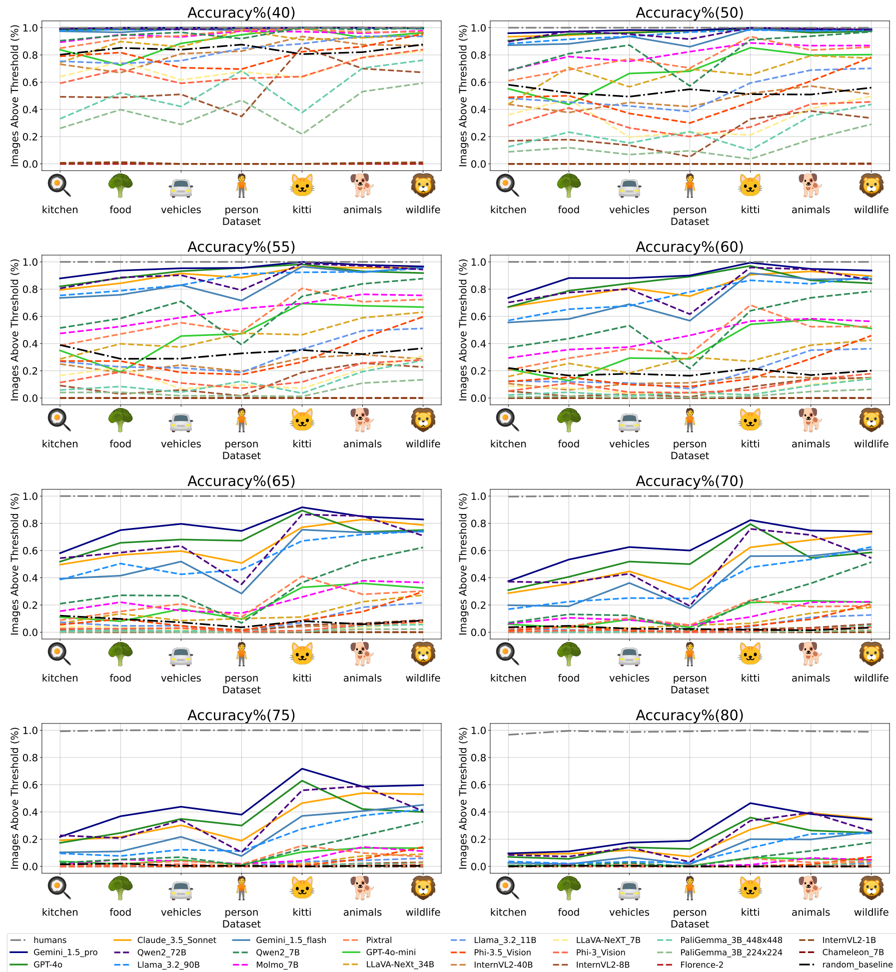


Figure 11: **Accuracy%(t) ↑thresholds for all models for each dataset**. Displayed as line plots, given the large number of models. Humans (dashed-dotted grey line) consistently achieve near-perfect scores of 1 across all displayed thresholds and datasets. Performance varies significantly across imaging domains, with dotted lines representing open-source models and solid lines indicating closed-source models.

C.3 Task Ranking 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, 25 = 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", (4), T2.1 "Object Occlusion Detection", and (5) 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 12, human baselines in Figure 13, all closed-source models in Figure 14, and all open-source models in Figure 15.

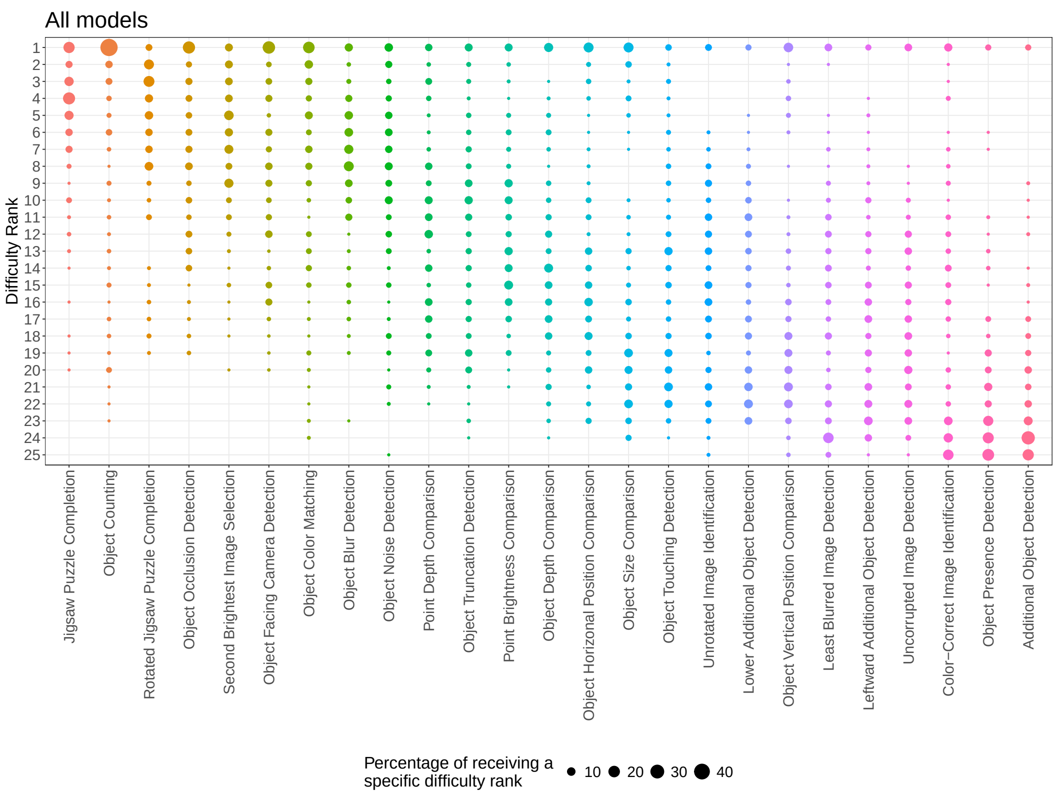


Figure 12: **All tested vision-language models, both open and closed-source, exhibit consistent patterns in task difficulty.** Aggregated ranking of tasks from easiest to most challenging, revealing systematic strengths and limitations shared across the complete set of evaluated models regardless of their accessibility.

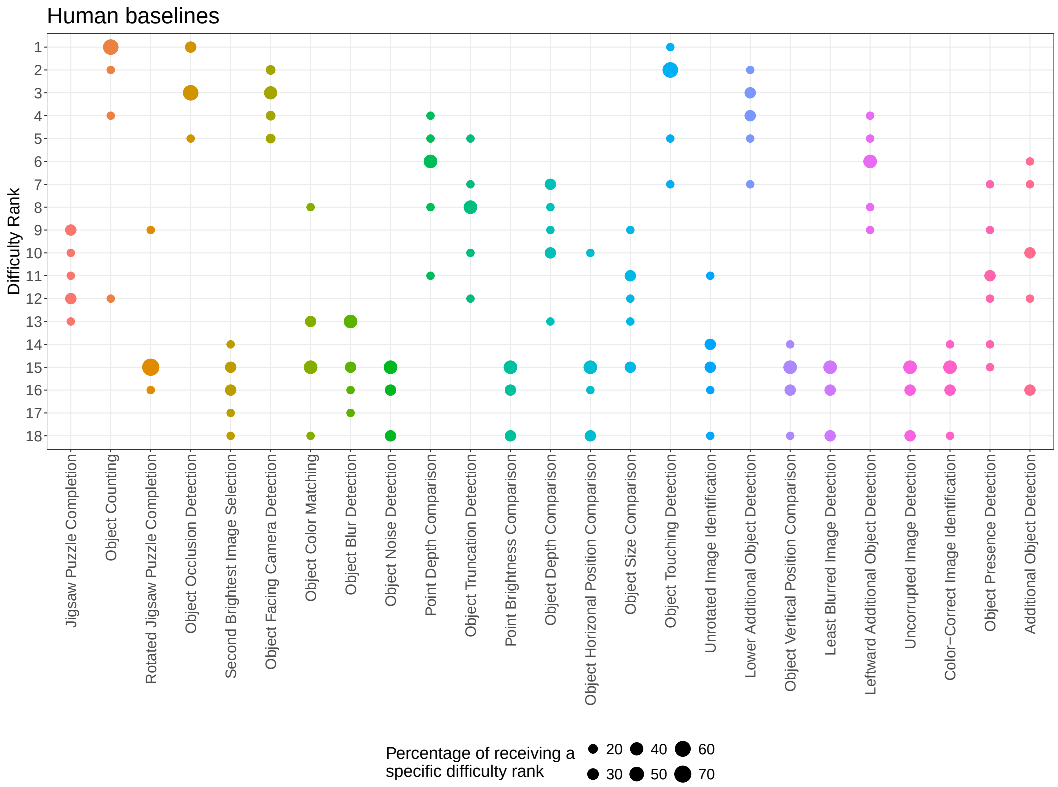


Figure 13: **Human performance reveals distinct patterns of task difficulty compared to models.** Ranking of vision tasks from easiest to most challenging based on human evaluator performance, providing a baseline for understanding natural visual capabilities.

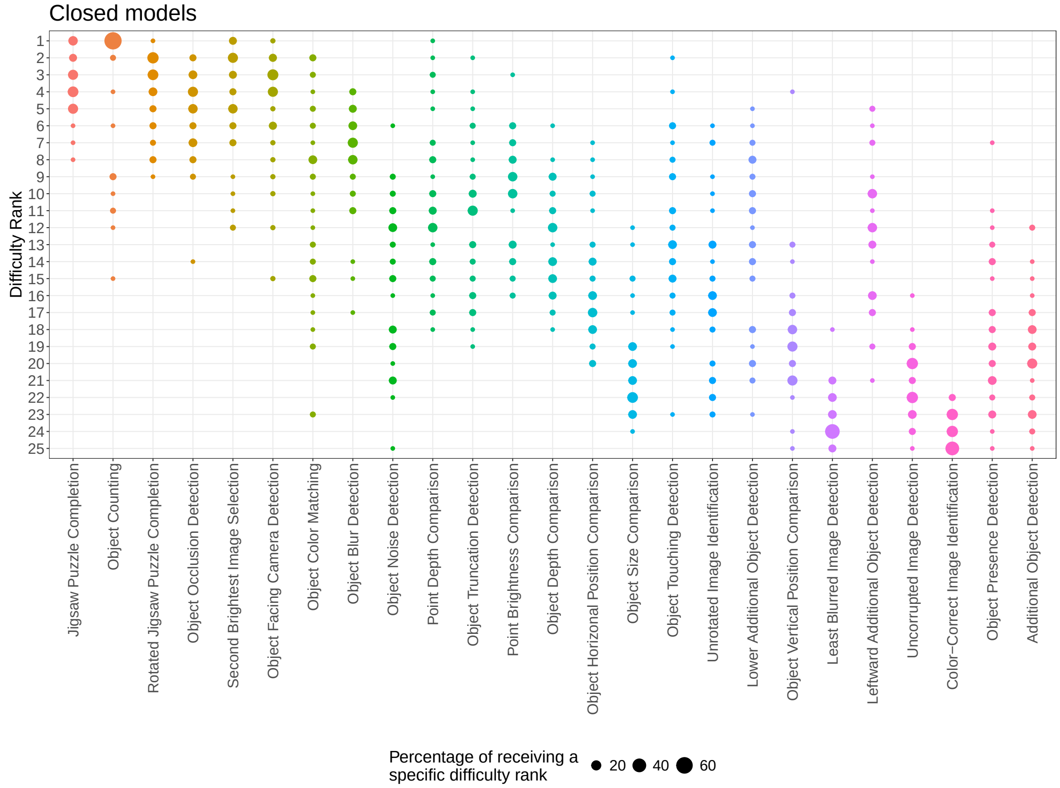


Figure 14: **Clsoed-sourced vision-language models demonstrate shared patterns in task difficulty despite architectural differences.** Aggregated ranking of tasks from easiest to most challenging, revealing systematic strengths and limitations common across all evaluated closed-source models.

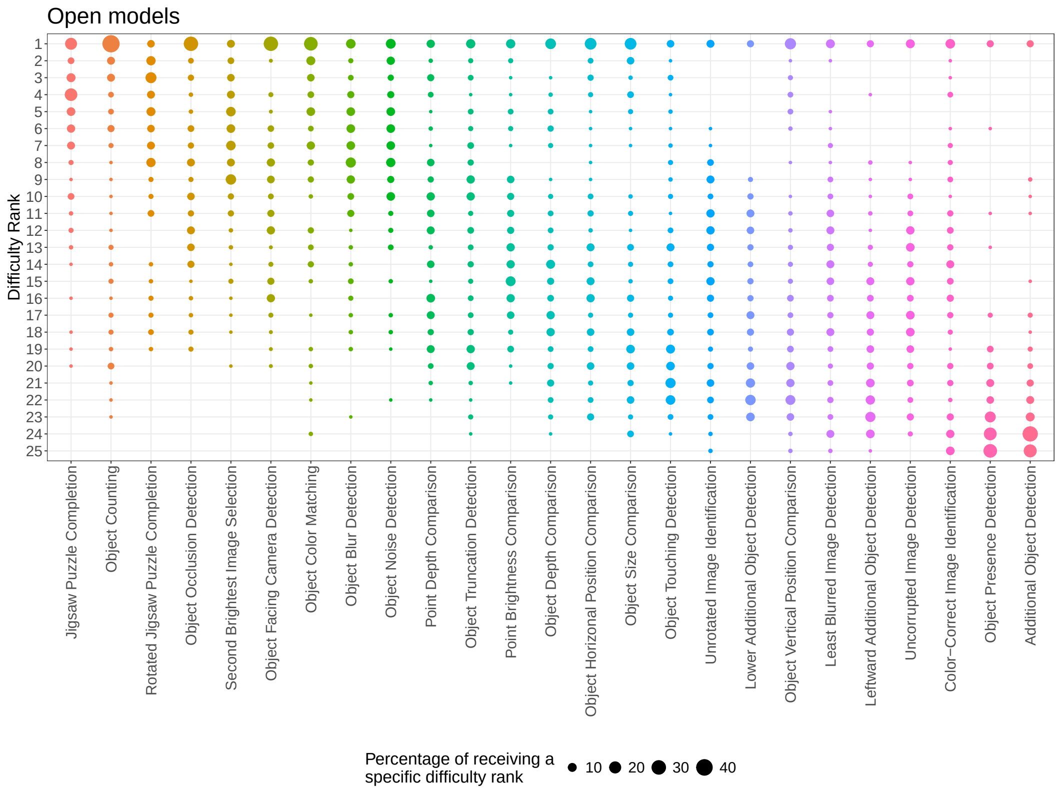


Figure 15: **Publicly available vision-language models exhibit consistent patterns in task difficulty across architectures.** Aggregated ranking of tasks from easiest to most challenging, revealing systematic strengths and limitations shared across all evaluated open-source models.

C.4 Accuracies Across Datasets

We display the regular Accuracy scores for all models per dataset in Table 9 and per task in Table 10.

Model	Overall	wildlife 🦁	animals 🐕	kitti 🐱	person 🧍	vehicles 🚘	food 🥦	kitchen 🍳
humans	93.66	93.43	93.87	94.60	95.15	92.41	93.58	92.59
Gemini_1.5_pro	72.44	74.58	75.35	77.96	70.84	71.74	69.98	66.60
GPT-4o	69.79	71.35	71.35	76.16	69.01	69.25	67.00	64.42
Claude_3.5_Sonnet	69.00	73.72	74.28	72.40	65.34	67.75	65.69	63.82
Qwen2_72B	68.76	70.79	75.00	74.63	61.70	68.23	66.22	64.74
Llama_3.2_90B	65.93	71.33	70.21	68.63	64.62	63.13	62.87	60.75
Gemini_1.5_flash	65.72	71.53	70.16	70.83	60.23	65.17	61.14	61.01
Qwen2_7B	59.71	67.82	65.01	62.22	51.49	59.32	57.34	54.80
Molmo_7B	57.89	61.11	61.57	59.77	57.09	56.12	56.37	53.19
Pixtral	56.80	59.77	59.17	63.20	54.44	56.08	54.03	50.90
GPT-4o-mini	54.90	59.19	59.42	60.80	53.63	53.71	46.90	50.66
LLaVA-NeXt_34B	53.26	57.47	56.67	53.57	54.26	50.72	53.11	47.02
Llama_3.2_11B	50.03	55.14	54.51	51.16	47.03	47.02	47.30	48.05
Phi-3.5_Vision	49.18	58.14	51.83	47.86	44.46	45.43	49.21	47.34
InternVL2-40B	48.14	49.48	49.98	50.64	47.37	47.98	44.69	46.85
LLaVA-NeXT_7B	44.52	48.84	46.44	41.38	42.39	41.52	47.27	43.79
Phi-3_Vision	44.39	48.99	47.41	42.43	41.62	42.12	45.97	42.18
InternVL2-8B	41.28	43.72	45.60	46.76	35.59	38.95	39.26	39.05
PaliGemma_3B_448x448	40.66	47.07	45.02	36.00	43.14	37.82	40.64	34.94
PaliGemma_3B_224x224	36.43	41.26	39.73	31.57	38.26	34.13	37.32	32.76
InternVL2-1B	16.69	18.05	15.51	18.31	15.94	14.54	19.58	14.89
Florence-2	15.47	17.80	16.02	15.01	18.29	13.85	16.67	10.65
Chameleon_7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 9: **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.

C.5 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.

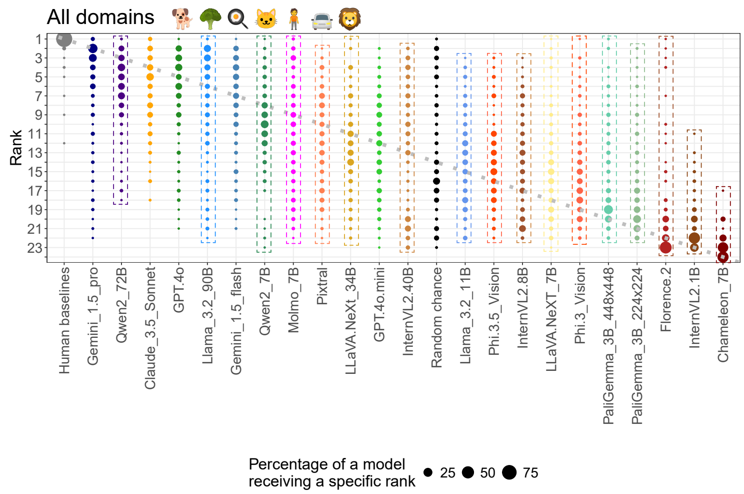


Figure 16: **Ranking variation across models for all datasets combined.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

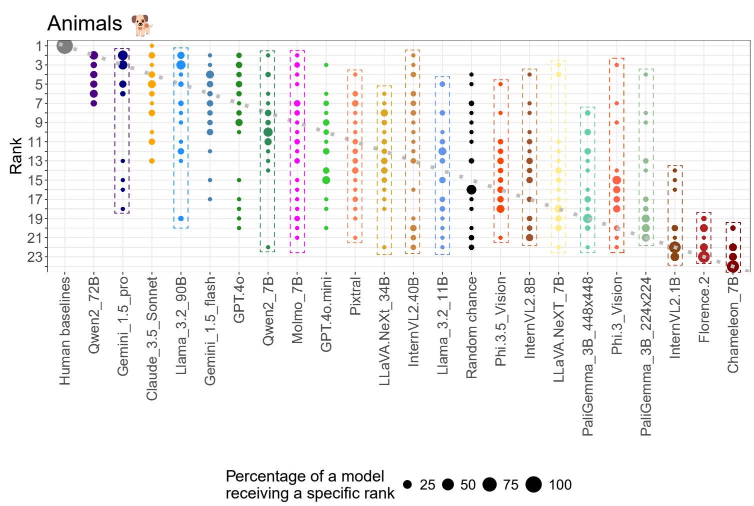


Figure 17: **Ranking variation across models for the animals dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

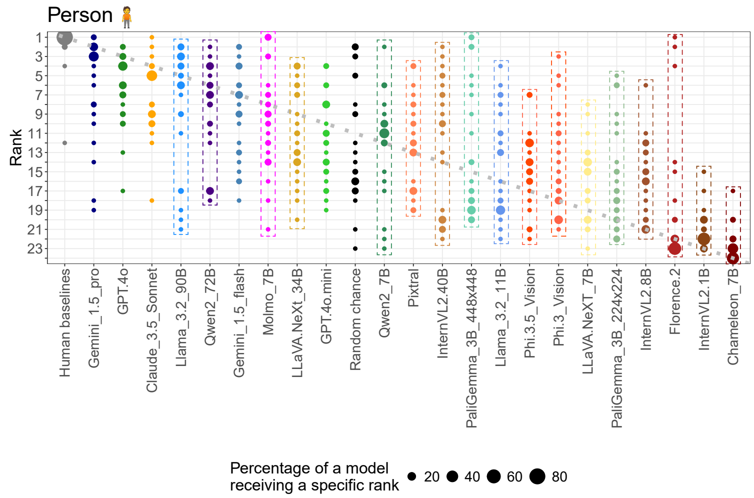


Figure 18: **Ranking variation across models for the person dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

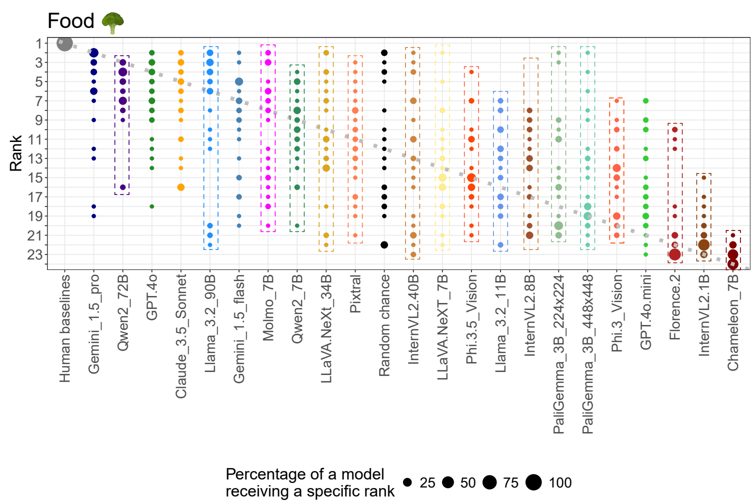


Figure 19: **Ranking variation across models for the food dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

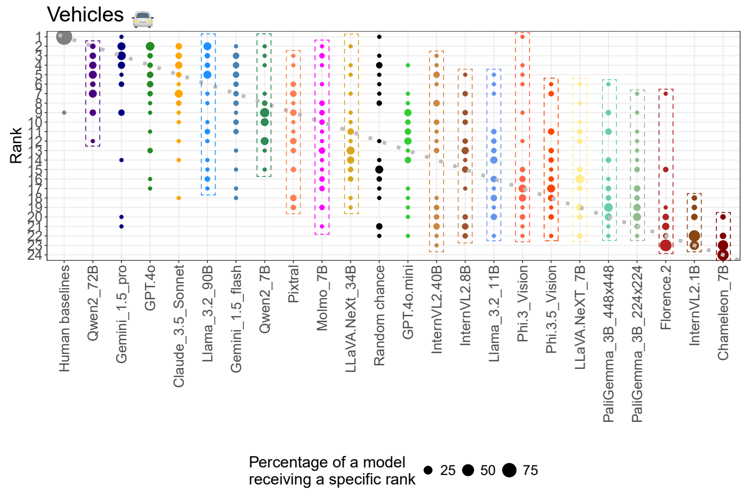


Figure 20: **Ranking variation across models for the vehicles dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

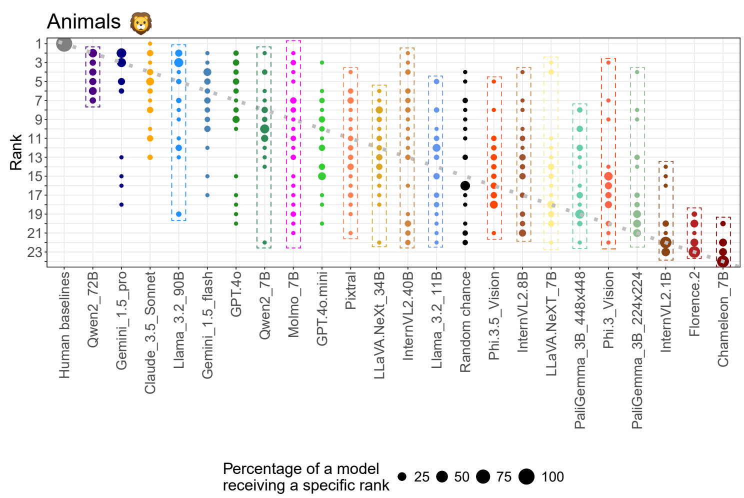


Figure 21: **Ranking variation across models for the wildlife dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

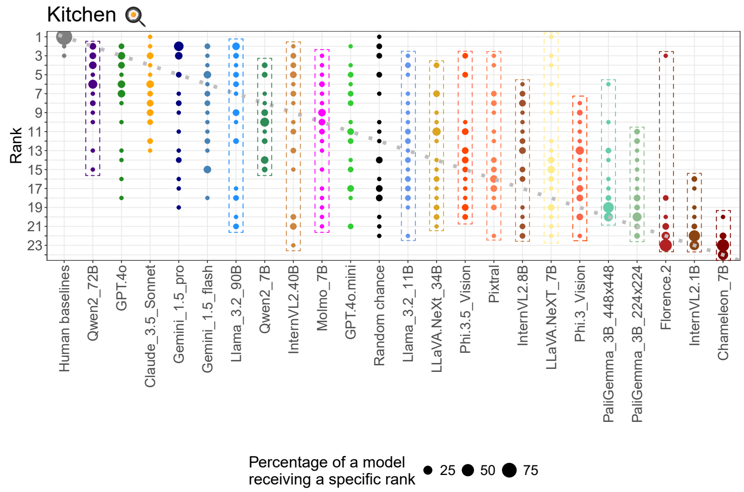


Figure 22: **Ranking variation across models for the kitchen dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

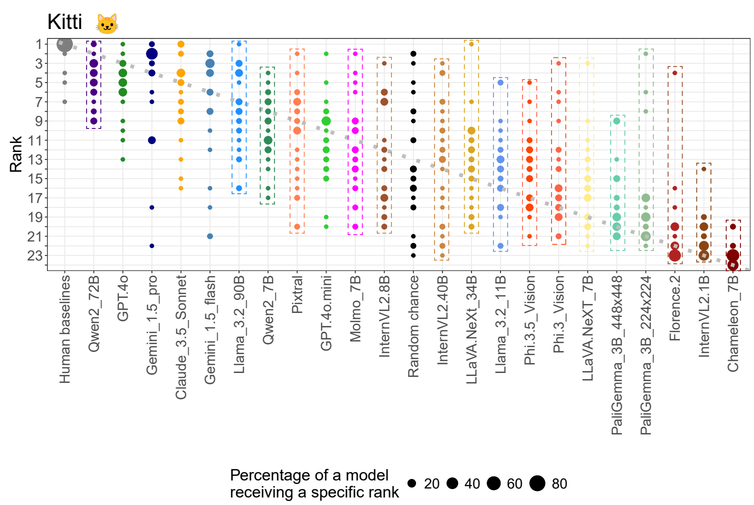


Figure 23: **Ranking variation across models for the kitti dataset.** Depicted as a scatter plot. The radius of each blob at position $(Model\_i, rank\_j)$ is proportional to the percentage that model $Model\_i$ achieved rank j. Open models are indicated by a dashed border.

C.6 Task Difficulty Comparison by Metadata Source and Ambiguity

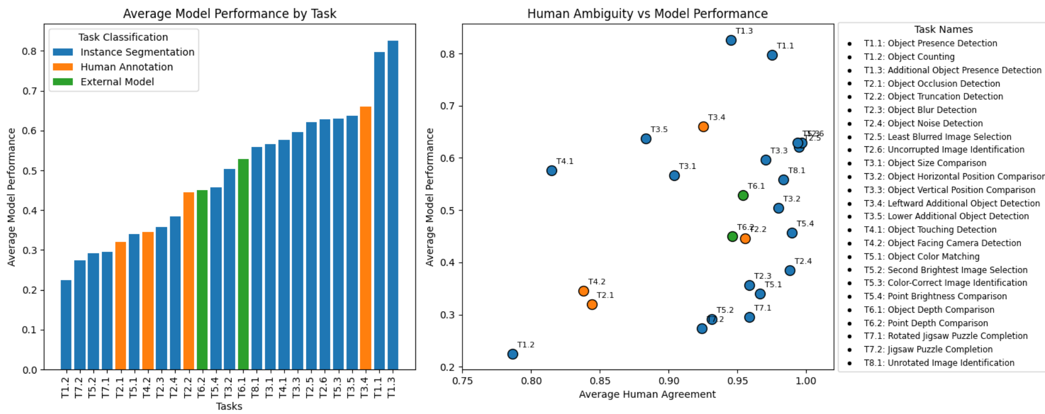


Figure 24: **Instance segmentations alone allow for the extraction of hard tasks.** (a) Tasks were classified in those extractable directly from instance segmentations (blue), requiring external models (green) and requiring human annotations (red). (b) Human ambiguity plotted against model performance.

C.7 Task Diversity: Model Accuracies Across Tasks

Here, we present the performance of models across all tasks, using Accuracy (see Table 10). Spider plots for closed-source, open-source, and varying model sizes are displayed in Figure 25.

Model	Rank	T1.1	T1.2	T1.3	T2.1	T2.2	T2.3	T2.4	T2.5	T2.6	T3.1	T3.2	T3.3	T3.4
humans		96.88	68.31	95.24	78.93	92.92	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	47.74	72.91	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	54.94	80.41	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	49.96	64.56	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	39.80	78.67	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	41.42	65.21	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	39.87	68.89	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	39.87	48.14	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	28.99	25.84	27.54	30.74	78.23	76.44	79.53	74.20	82.49	67.28
Pixtral	9	86.84	7.14	88.75	43.64	37.04	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	33.83	39.76	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	44.10	25.49	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	42.20	72.40	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	43.93	31.32	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	0.00	39.81	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	43.45	30.54	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	41.00	27.84	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	0.06	61.33	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	31.74	47.41	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	35.85	42.85	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	0.00	14.20	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	28.27	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	0.00	0.00

Model	Rank	T3.5	T4.1	T4.2	T5.1	T5.2	T5.3	T5.4	T6.1	T6.2	T7.1	T7.2	T8.1
humans		82.57	75.34	82.39	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	40.97	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	51.08	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	47.05	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	53.98	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	58.78	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	49.15	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	39.00	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	56.05	28.79	35.34	69.54	67.60	78.68	53.21	28.87	26.05	84.32
Pixtral	9	63.37	70.13	36.51	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	45.14	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	46.67	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	36.83	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	40.56	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	0.00	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	35.97	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	47.50	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	0.00	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	47.86	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	31.76	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	0.00	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	0.00	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	0.00

Table 10: **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.

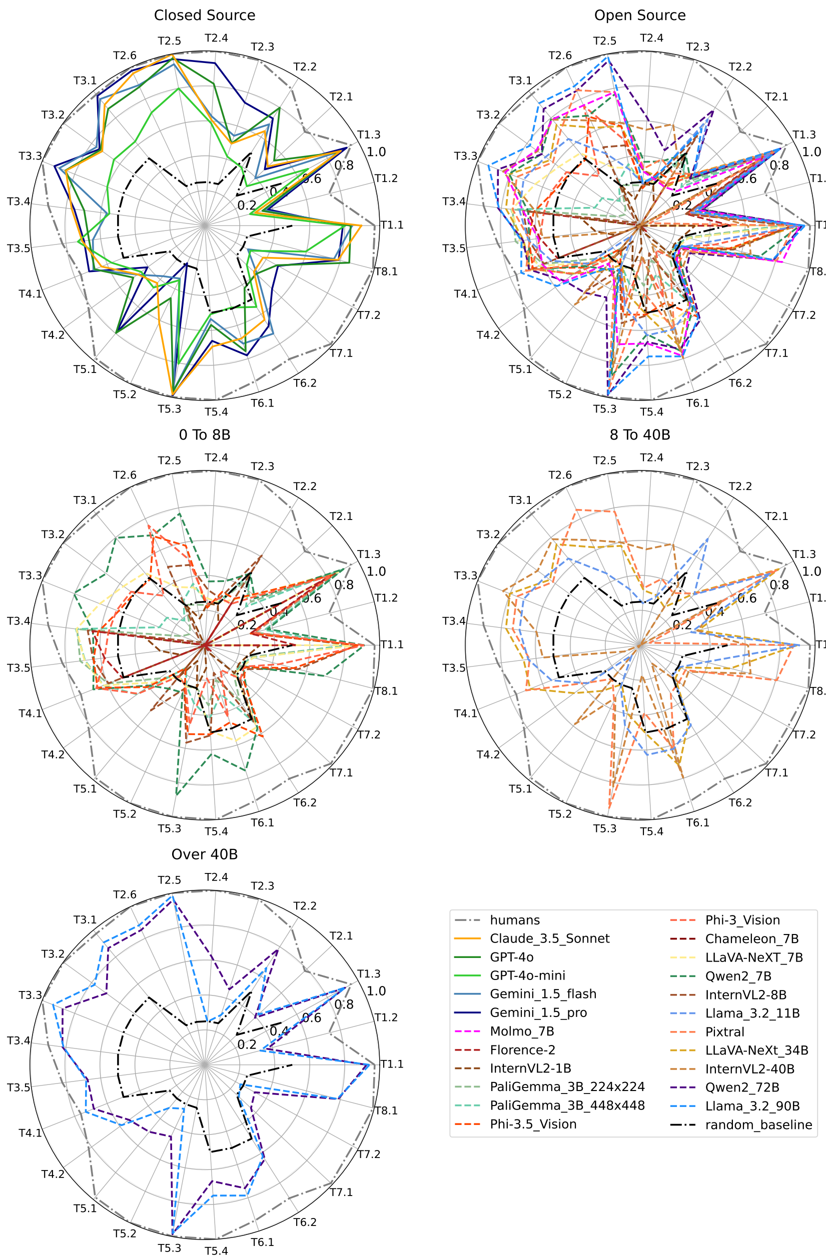


Figure 25: **Model architecture and size significantly impact performance patterns across diverse vision tasks.** Spider plots reveal distinct performance profiles between open-source (dashed lines) and closed-source (solid lines) models across our comprehensive task framework. Each axis represents task-specific accuracy, demonstrating how different model characteristics influence capabilities.

C.8 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 26, for all 22 models in Figure 27, and Area under the Accuracy%(t) Curves in Table 11.

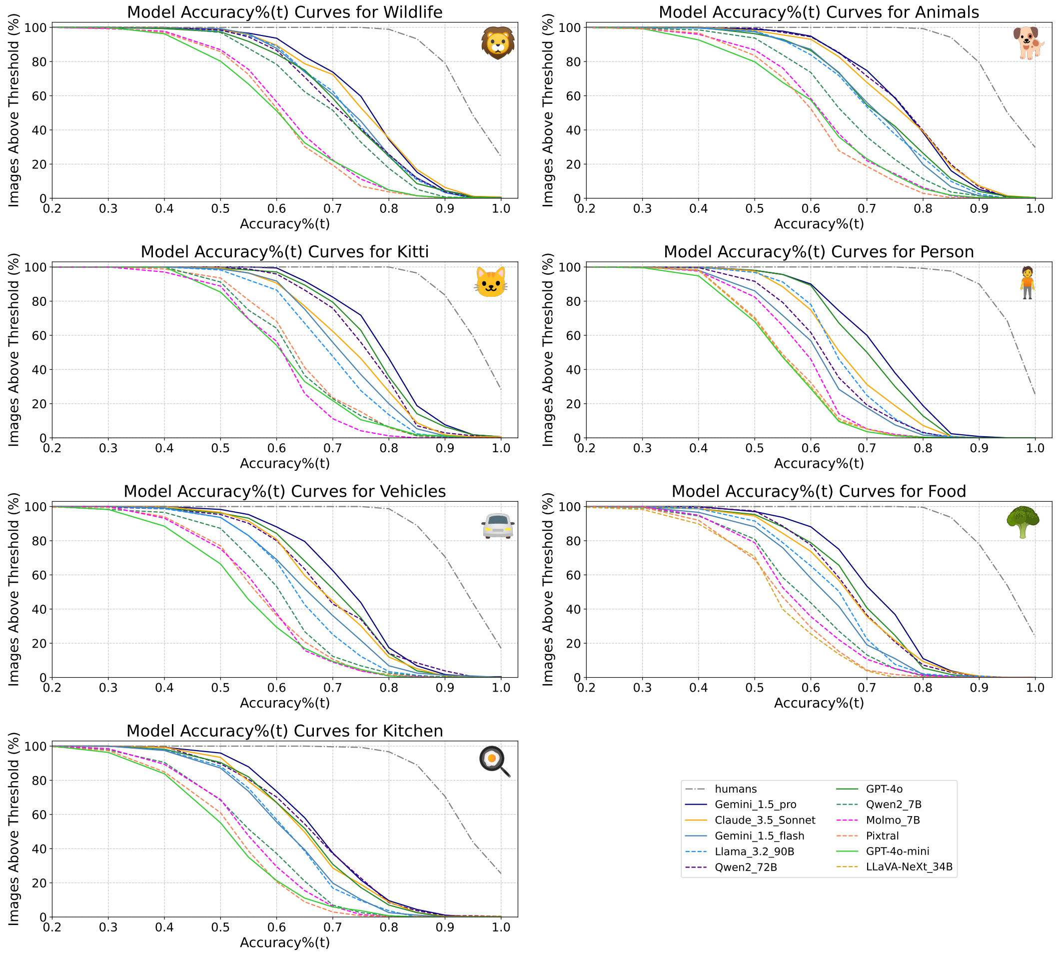


Figure 26: **Accuracy% (t) curves for the top 10 models across each dataset,** with a maximum score of 1. Humans (dashed-dotted grey line) consistently achieve the highest performance. Dashed lines indicate open-source models, while solid lines represent closed-source models. The area under the Accuracy% (t) curves, detailed in Table 11, highlights significant variations in model rankings across domain-specific datasets for the same tasks.

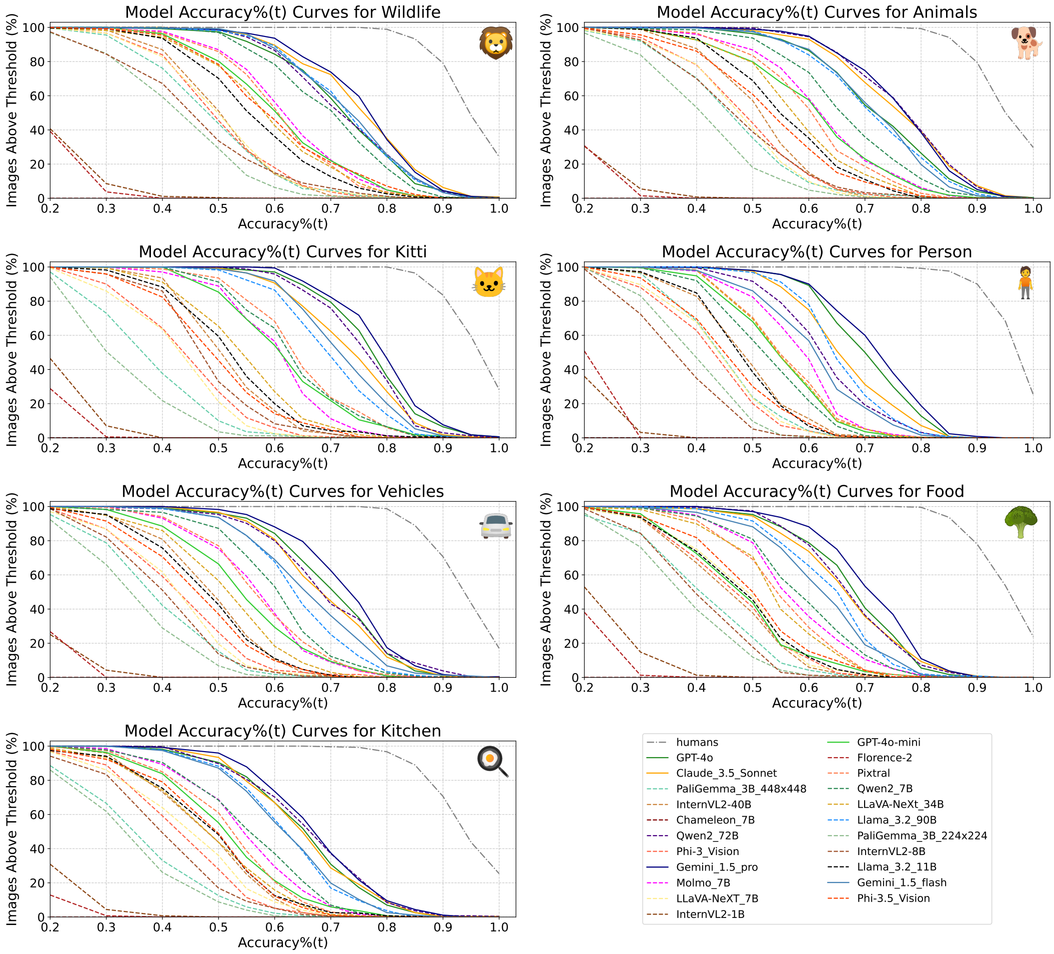


Figure 27: **Accuracy% (t) curves for all models across each dataset,** with a maximum score of 1. Humans (dashed-dotted grey line) consistently achieve the highest performance. Dashed lines indicate open-source models, while solid lines represent closed-source models. The area under the Accuracy% (t) curves, detailed in Table 11, highlights significant variations in model rankings across domain-specific datasets for the same tasks.

Model	Overall	animals 🐕	food 🥦	kitchen 🍳	kitti 🐱	person 🧍	vehicles 🚘	wildlife 🦁
humans	74.28	74.40	74.34	73.08	75.18	75.89	73.01	74.10
Gemini_1.5_pro	52.95	56.02	50.39	46.81	58.51	51.35	52.17	55.42
GPT-4o	50.19	51.87	47.24	44.52	56.74	49.59	49.72	51.64
Claude_3.5_Sonnet	49.80	55.36	46.24	44.73	53.03	45.92	48.48	54.85
Qwen2_72B	49.57	56.15	46.89	45.63	55.38	42.27	48.85	51.82
Llama_3.2_90B	46.57	51.07	43.16	41.45	49.25	45.00	43.66	52.42
Gemini_1.5_flash	46.33	51.10	41.61	41.39	51.56	40.45	45.60	52.63
Qwen2_7B	40.35	46.08	37.87	34.94	42.81	31.93	39.73	49.11
Molmo_7B	38.27	41.84	36.78	33.96	39.78	37.60	36.29	41.65
Pixtral	37.34	40.00	34.23	31.42	43.69	35.06	36.56	40.45
GPT-4o-mini	35.50	40.49	27.34	31.03	41.31	34.08	34.00	40.29
LLaVA-NeXt_34B	33.65	37.26	33.23	27.67	33.63	34.72	30.71	38.28
Llama_3.2_11B	30.44	35.35	27.17	28.06	31.75	27.44	27.19	36.10
random_chance	30.24	29.39	30.31	30.86	29.78	30.39	29.70	31.27
Phi-3.5_Vision	29.87	32.93	29.22	28.01	28.96	24.93	25.67	39.35
InternVL2-40B	28.48	30.60	25.16	27.14	30.76	27.85	27.89	29.95
LLaVA-NeXT_7B	24.94	26.88	27.10	24.04	22.03	23.12	21.81	29.64
Phi-3_Vision	24.89	27.88	26.05	22.74	23.41	22.12	22.60	29.44
InternVL2-8B	22.21	26.79	19.83	20.09	27.41	16.17	19.81	25.37
PaliGemma_3B_448x448	21.44	25.70	20.86	15.81	17.04	24.13	18.40	28.16
PaliGemma_3B_224x224	17.28	20.82	17.61	13.99	12.35	18.80	14.72	22.65
InternVL2-1B	2.62	2.14	4.26	2.06	3.03	2.12	1.68	3.03
Florence-2	1.74	1.68	2.03	0.72	1.50	2.54	1.34	2.33
Chameleon_7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 11: **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.

C.9 Task Correlation

Here we present the correlation between tasks, quantified in the heatmap in Figure 28 and visualized in the dendrogram in Figure 29. The heatmap shows pairwise task correlations, while the dendrogram highlights clusters of tasks with similar performance patterns across models

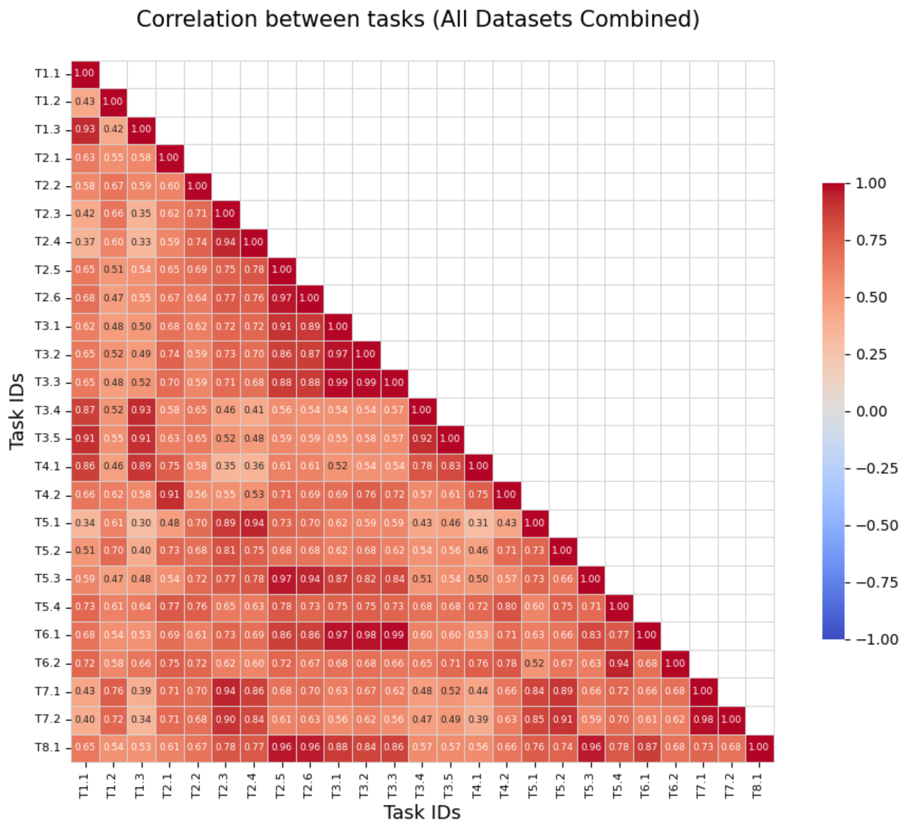


Figure 28: **Task performance correlations reveals insights on related visual capabilities.** Heatmap visualizing pairwise correlations between task performances across all datasets, with values ranging from -1 (anti-correlated) to 1 (highly correlated).

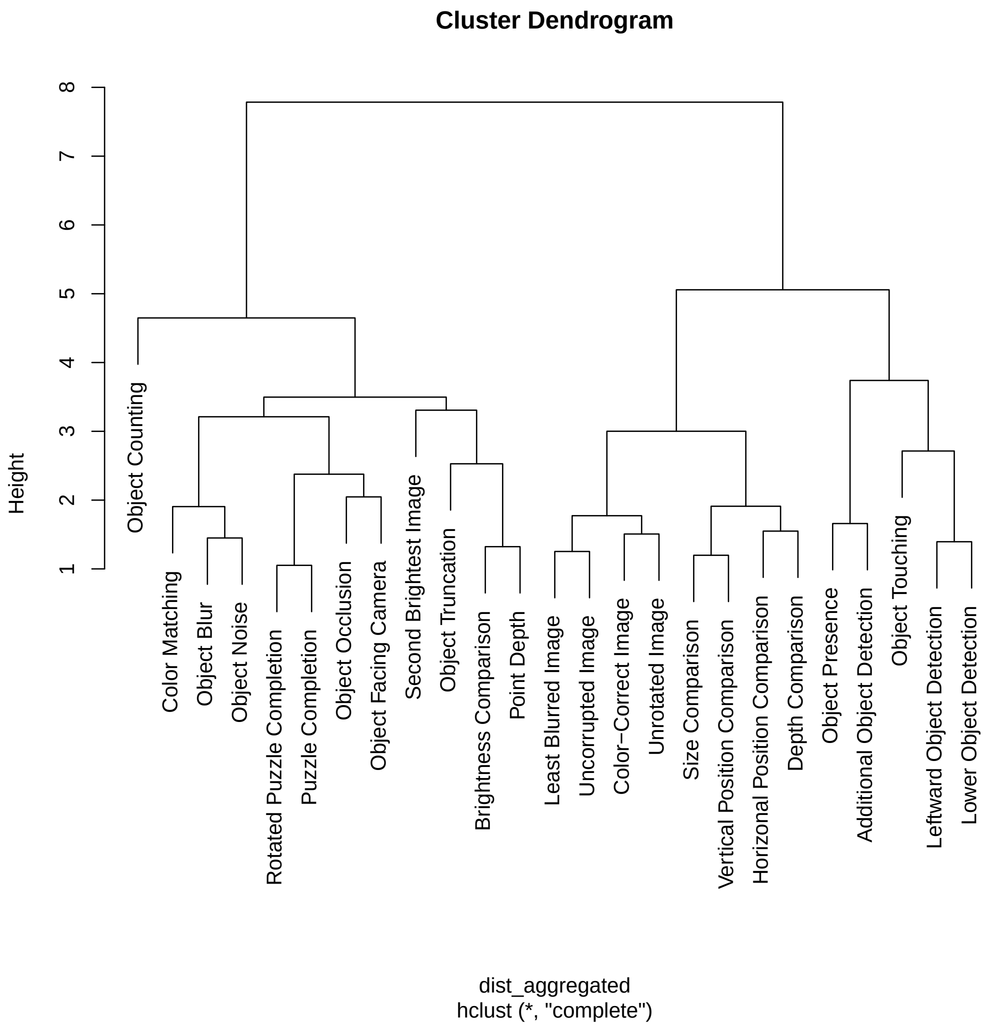


Figure 29: **Hierarchical clustering reveals natural groupings of visually related tasks across domains.** Dendrogram visualization of task relationships based on model performance across seven domains, confirming and extending the correlation patterns observed in the heatmap analysis.