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.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL https://arxiv.org/abs/2303.12712.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024a. URL https://arxiv.org/abs/2403.20330.

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*, 2024b.

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.

Florian E. Dorner and Moritz Hardt. Don't label twice: Quantity beats quality when comparing binary classifiers on a budget. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR, July 2024. URL https://proceedings.mlr.press/v235/dorner24a.html.

Anca Dumitrache, Lora Aroyo, and Chris Welty. Achieving expert-level annotation quality with crowdtruth: The case of medical relation extraction. In *BDM2I@ISWC*, 2015. URL https://api.semanticscholar.org/CorpusID:10208514.

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.

Leslie Guzene, Arnaud Beddok, Christophe Nioche, Romain Modzelewski, Cedric Loiseau, Julia Salleron, and Juliette Thariat. Assessing interobserver variability in the delineation of structures in radiation oncology: A systematic review. *International Journal of Radiation Oncology*Biology*Physics*, 115(5):1047–1060, 2023. ISSN 0360-3016. doi: 10.1016/j.ijrobp.2022.11.021.

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.

Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, 8(4):614–629, 2020.

Vasily Kostumov, Bulat Nutfullin, Oleg Pilipenko, and Eugene Ilyushin. Uncertainty-aware evaluation for vision-language models. *arXiv preprint arXiv:2402.14418*, 2024.

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*, 2024a.

Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, et al. Ii-bench: An image implication understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2406.05862*, 2024b.

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, 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, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102306. URL https://www.sciencedirect.com/science/article/pii/S1361841521003510.

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.

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.

Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. *arXiv preprint arXiv:2406.14544*, 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.

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. *arXiv preprint arXiv:2407.17596*, 2024.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 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.

Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, and Reinhard Koch. Is one annotation enough? a data-centric image classification benchmark for noisy and ambiguous label estimation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

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, 2024. URL https://arxiv.org/abs/2406.16860.

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 2018.

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.

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*, 2024b.

A EXAMPLE TASKS FOR AN IMAGE

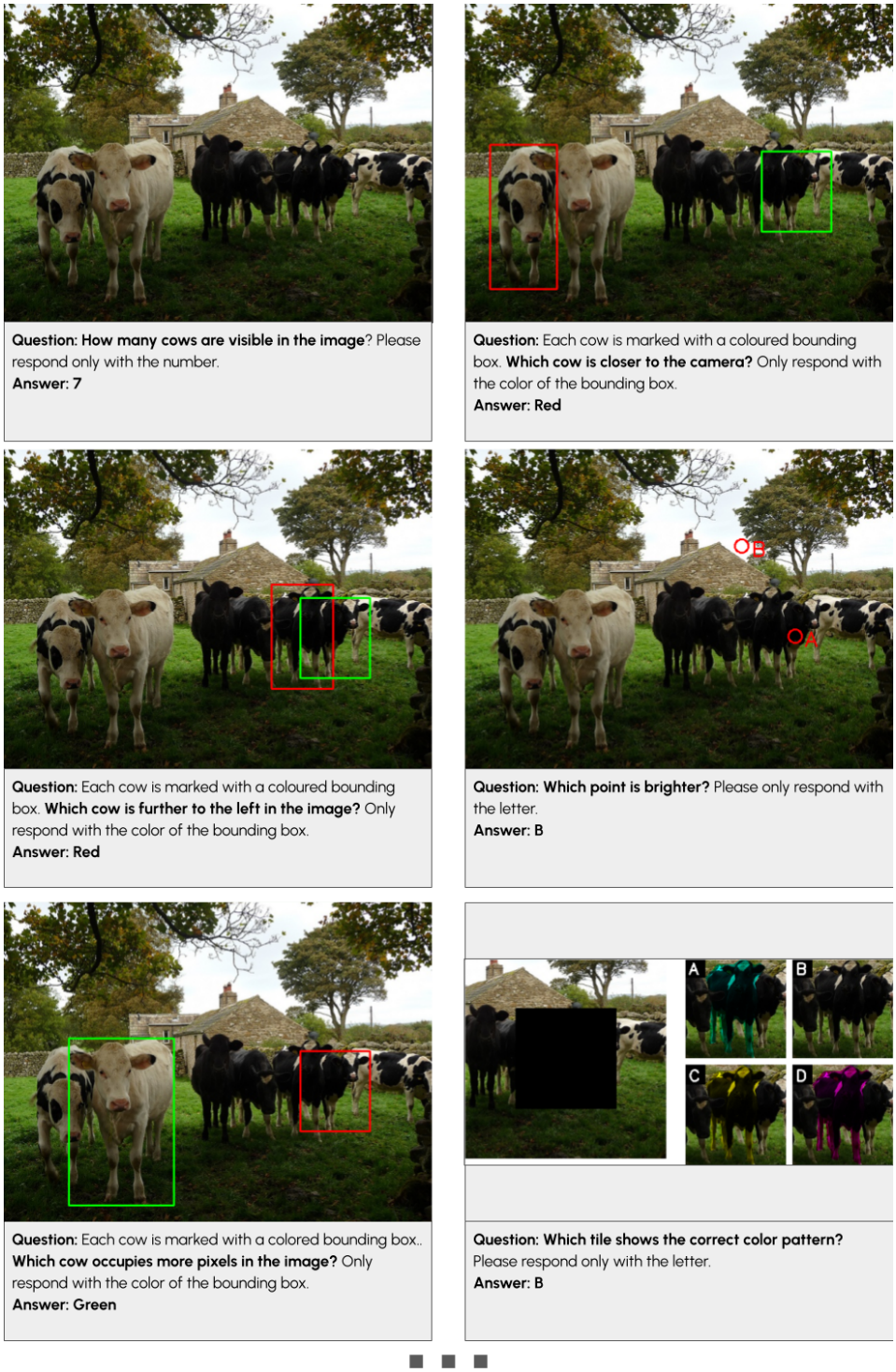

Figure 6: **For each image, we generate numerous tasks covering diverse perception abilities.**

B EXAMPLE OF HUMAN METADATA ANNOTATION

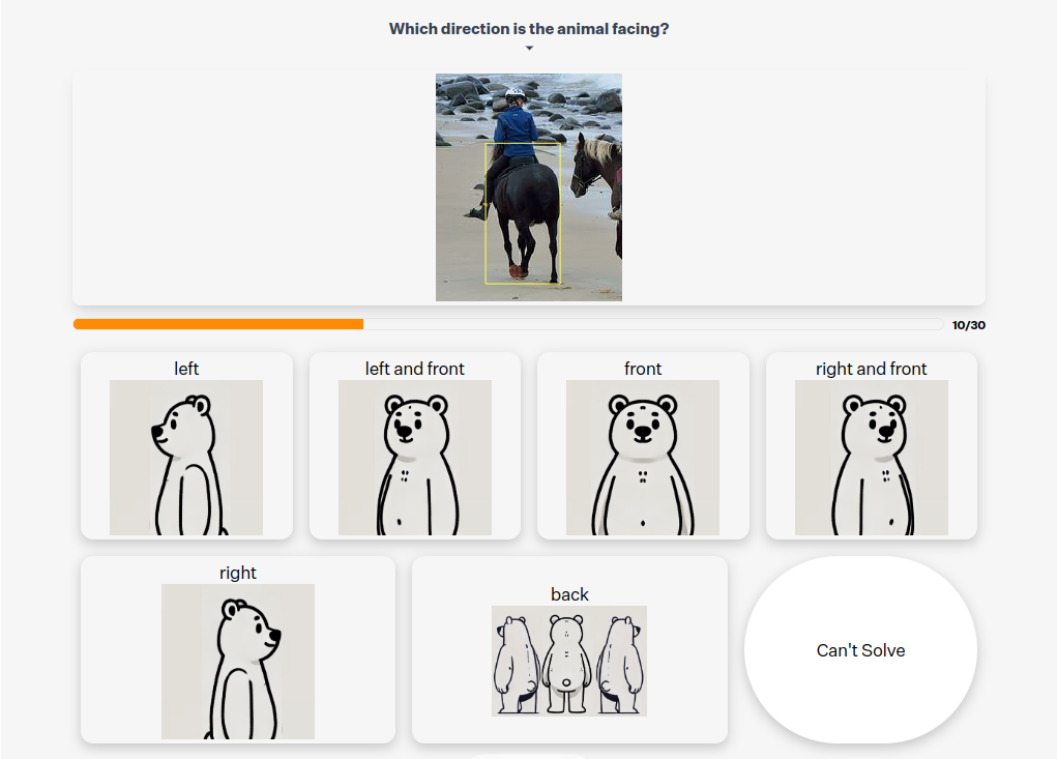

Figure 7: **Exemplary initial human metadata enrichment task.** These annotations were used to enrich the objects with human generated metadata.

C CVPR 2024 PAPER ANALYSIS

Table 2: CVPR 2024 paper analysis summary.

CVPR 2024	
Total number of papers	2,708
With New or modified dataset:	397
Without new or modified dataset:	2,311

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.

D ACCURACY%(T) CURVES ACROSS DATASETS

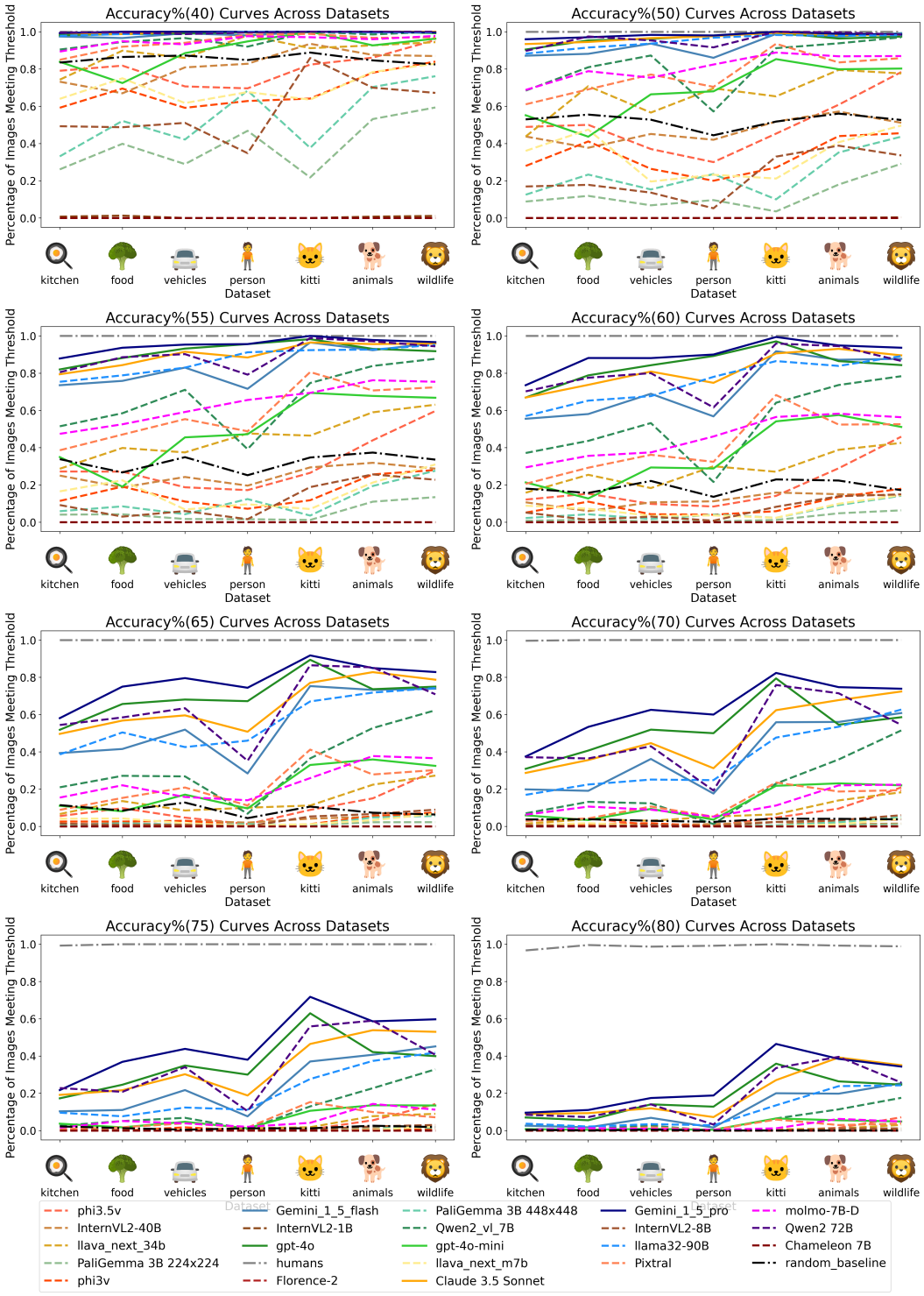

Figure 8: **The need for specific in-domain evaluation is demonstrated by the high performance variability across imaging domains.** The Accuracy%(t) metric represents the percentage of images for which at least a specified proportion of questions are correctly answered by the mode.

E MODEL OVERVIEW

Table 3: VLM Benchmark Models used in this study

Accessibility	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	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

F OVERVIEW OF VLM BENCHMARK ANNOTATION PROCESSES

Table 4: Overview of VLM Benchmark Annotation Processes

Benchmark	Annotators reported?	Raters per image	Comment
Blink	Yes	2 per image	Two annotators (co-authors) assigned per task. Exception: one question type received single annotation.
MMBench	No	N/A	Volunteers (students) expanded initial question set.
MME	No	N/A	Number of annotators unclear
MMStar	Yes	N/A	Three experts reviewed. Unclear if all samples seen by all.
MM-Vet v2	No	N/A	GPT-4V generated drafts, experts reviewed. Exact number undisclosed.
MMT-Bench	Yes	50 in total	”Dozens of co-authors” and 50 students assisted.
WildVision	Yes	1 per image	Crowdsourced. Cohen’s Kappa: 0.59.
MMMU	Yes	N/A	50 annotators, college students from diverse disciplines.
II-Bench	Yes	N/A	50 students collected and annotated images.
Vibe-Eval	Yes	N/A	22 group members collected prompts.
TouchStone	No	N/A	Manually annotated, no statistical info provided.
Seed-Bench-2	No	N/A	Partly manually annotated, number not given.
MME-RealWorld	Yes	N/A	25 professional annotators, 7 MLLM experts. Task distribution unclear.

G THREE METADATA SOURCES

Table 5: Metadata sources used for enriching instance segmentation datasets.

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

H MODEL ACCURACY%(T) CURVES FOR EACH DATASET

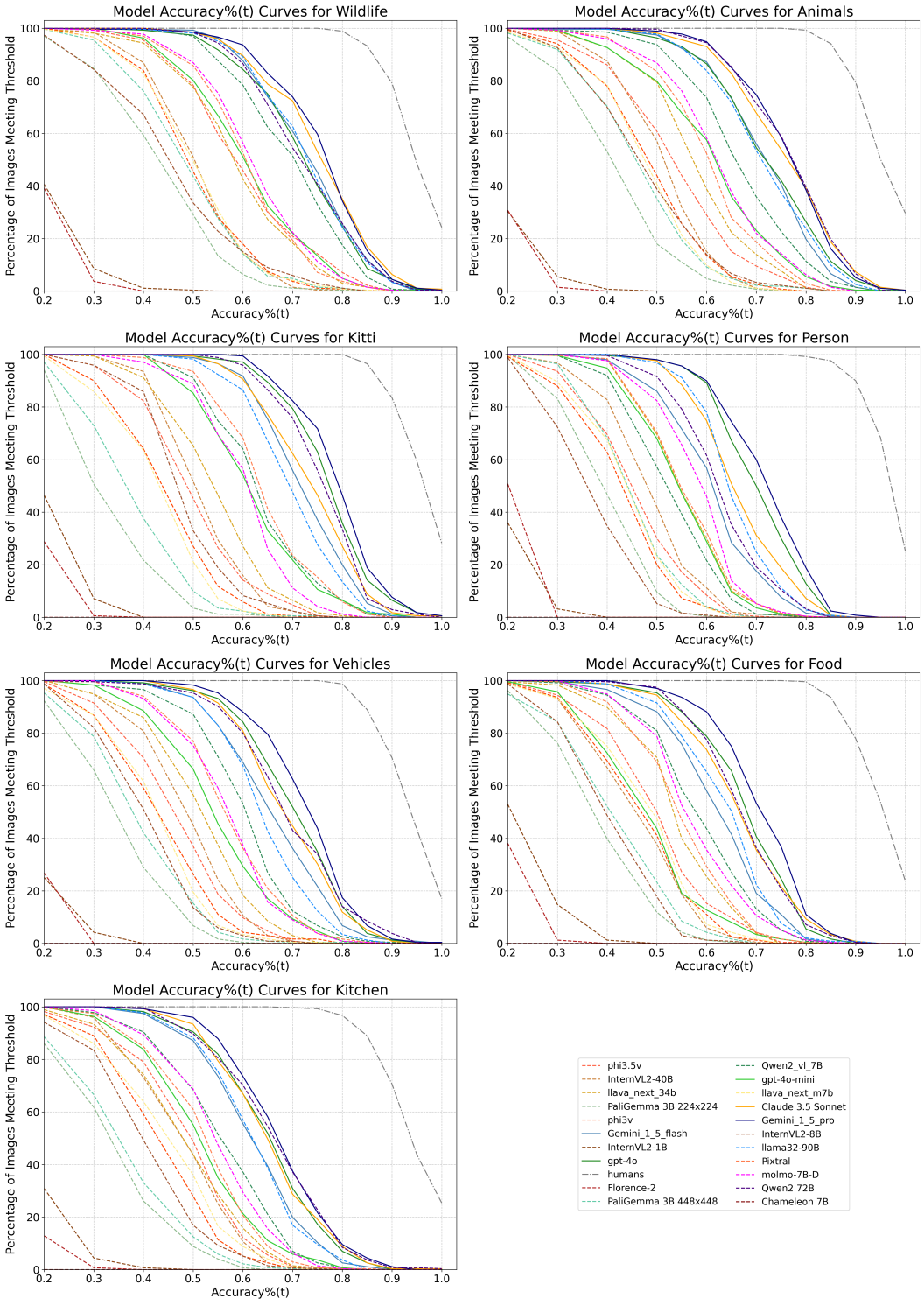

Figure 9: **Model Accuracy%(t) curves per dataset.** We observe performance variability across imaging domains and across models. Exemplary for animals, the best open model Qwen 72B (purple dashed) is almost on par with the best closed Model Gemini 1.5 pro (solid blue).

I OVERVIEW VLM TASKS

Table 6: Overview of VLM Benchmark Tasks

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)

J EXTRACTING HARD TASKS FROM INSTANCE SEGMENTATIONS

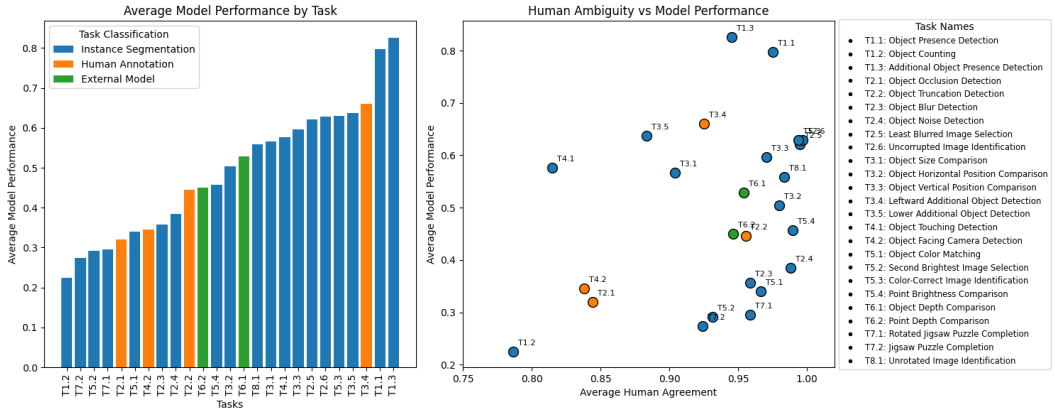

Figure 10: **Instance segmentations alone allow for the extraction of hard tasks.** (a) Tasks were classified in those extractable directly from instance segmentations (blue), requiring external models (green) and requiring human annotations (red). (b) Human ambiguity plotted against model performance.

K RANKING COMPARISON BETWEEN MODELS AND HUMANS

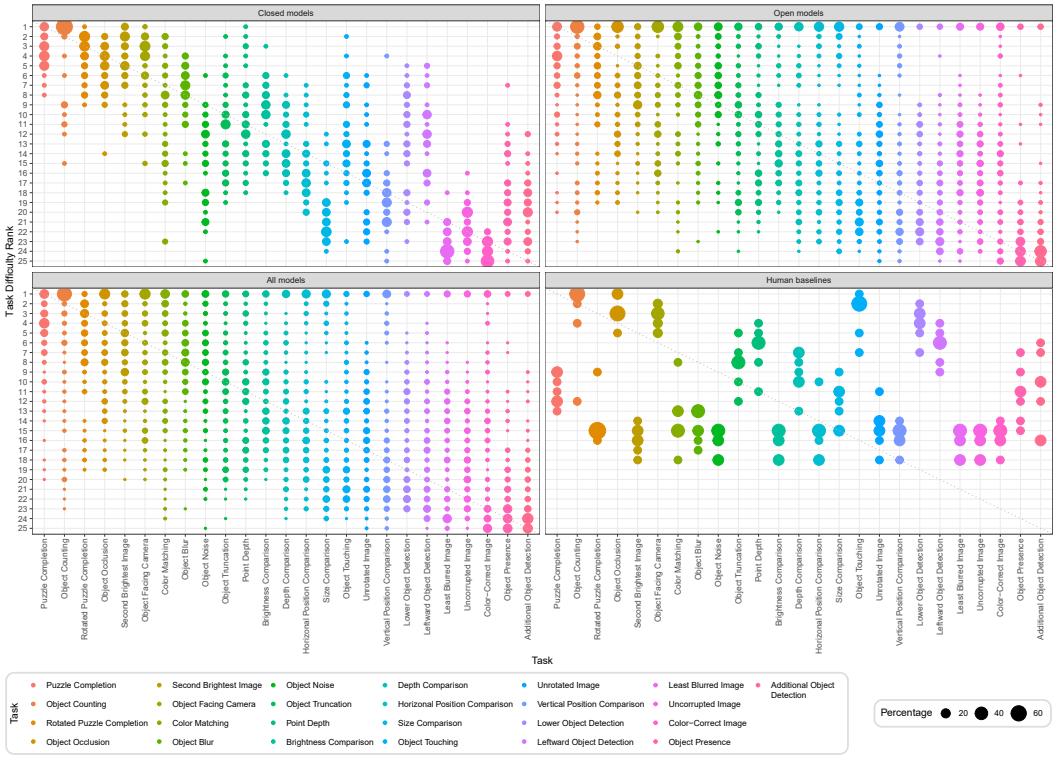

Figure 11: **Task ranking differs between models and human raters.** The plot shows the difficulty of tasks based on aggregated model scores (1 = hardest task, 25 = easiest task). The radius of the blob indicates how often a task was assigned a difficulty rank when considering all seven domains and all 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".