UNSOLVABLE PROBLEM DETECTION FOR VISION LANGUAGE MODELS

Atsuyuki Miyai[1] **Jingkang Yang**[2] **Jingyang Zhang**[3] **Yifei Ming**[4]
Qing Yu[1,5] **Go Irie**[6] **Yixuan Li**[4] **Hai Li**[3] **Ziwei Liu**[2] **Kiyoharu Aizawa**[1]
[1]The University of Tokyo [2]S-Lab, Nanyang Technological University
[3]Duke University [4]University of Wisconsin-Madison [5]LY Corporation [6]Tokyo University of Science
https://github.com/AtsuMiyai/UPD/

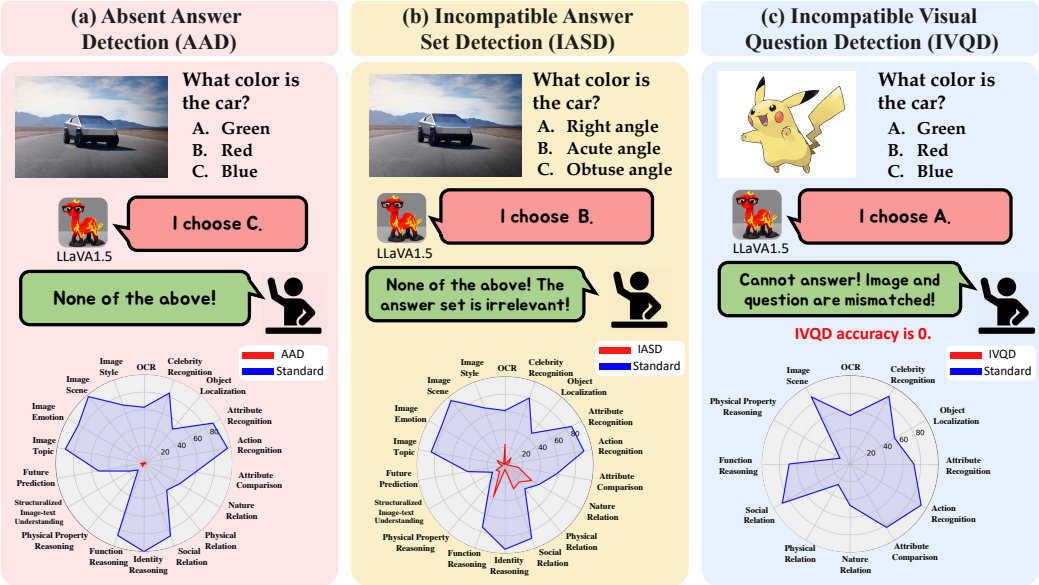


Figure 1: **The Unsolvable Problem Detection (UPD) Challenges**. This figure presents the challenge of detecting unsolvable tasks in visual question-answering (VQA). Current vision-language models (VLMs), such as LLaVA-1.5, show adequate performance (blue) on standard VQA tasks in MMBench. However, they exhibit a notable deficiency (red) in appropriately refraining from answering unsolvable VQA problems.

ABSTRACT

This paper introduces a novel and significant challenge for Vision Language Models (VLMs), termed **Unsolvable Problem Detection (UPD)**. UPD examines the VLM's ability to withhold answers when faced with unsolvable problems in the context of Visual Question Answering (VQA) tasks. UPD encompasses three distinct settings: Absent Answer Detection (AAD), Incompatible Answer Set Detection (IASD), and Incompatible Visual Question Detection (IVQD). To deeply investigate the UPD problem, extensive experiments indicate that most VLMs, including GPT-4V and LLaVA-NeXT-34B, struggle with our benchmarks to varying extents, highlighting significant room for the improvements. To address UPD, we explore both training-free and training-based solutions, offering new insights into their effectiveness and limitations. We hope our insights, together with future efforts within the proposed UPD settings, will enhance the broader understanding and development of more practical and reliable VLMs. A longer version of this paper is available at https://arxiv.org/abs/2403.20331.

1 INTRODUCTION

Imagine a diligent student sitting for a history examination, meticulously working through the questions. Suddenly, they encounter a peculiar query: "From which archaeological site was this artifact unearthed?" However, instead of an image of an ancient artifact, the provided picture is unexpectedly of Pikachu. Confounded yet astute, the student raises a hand and remarks, "Teacher, there

seems to be a mistake with this question!" This scenario, illustrating the human ease of identifying an unsolvable or out-of-place exam question, highlights a critically demanding challenge for Vision Language Models (VLMs).

In recent years, following the revolutionary development of Large Language Models (LLMs) (Chen et al., 2023; vic, 2023; Li et al., 2023d; Touvron et al., 2023a; Wei et al., 2023; Zhao et al., 2023b), VLMs (Awadalla et al., 2023; Bubeck et al., 2023; Dai et al., 2023; Touvron et al., 2023a; Wang et al., 2023d; Ye et al., 2023; Zhu et al., 2024; Li et al., 2023a; Lin et al., 2024) have also demonstrated profound capabilities in various applications and significantly enhance the performance in image reasoning tasks (Antol et al., 2015; Liu et al., 2024b; 2023c; Yue et al., 2024). However, the reliability of these models, especially in providing accurate and trustworthy information, has become a growing concern. Often, these models produce incorrect or misleading information, a phenomenon known as "hallucination", highlighting the importance of safety in large models (Bommasani et al., 2021; Wang et al., 2023a; Mallen et al., 2023; Zhang et al., 2023b; Huang et al., 2023a; Sun et al., 2024; Lu et al., 2024a).

We focus on a specific aspect of the VLM trustworthiness: the challenge of identifying out-of-place or unsolvable questions, analogous to the student's dilemma in our opening scenario. We introduce **Unsolvable Problem Detection (UPD)** for VLMs, a novel task that assesses a model's ability to withhold an answer when confronted with unsolvable problems. UPD is explored through three distinct settings: Absent Answer Detection (AAD), Incompatible Answer Set Detection (IASD), and Incompatible Visual Question Detection (IVQD). These settings are designed to evaluate the model's proficiency in managing unanswerable queries due to various forms of irrelevance and discrepancies. The illustration for each setting is shown in Fig. 1 and will be explained in Sec. 2.

We carefully adapt MMBench (Liu et al., 2023c), a VLM benchmark consisting of single-choice questions covering different ability dimensions (*e.g.,* object localization and social reasoning), to create the three benchmarks for our UPD settings: MM-AAD Bench, MM-IASD Bench, and MM-IVQD Bench. We evaluate five recent powerful open-source VLMs including LLaVA-1.5-13B (Chung et al., 2022; Liu et al., 2023a), CogVLM-17B (Wang et al., 2023d), Qwen-VL-Chat (Bai et al., 2023), the more recent LLaVA-NeXT (13B, 34B) (Liu et al., 2024c), and two close-source VLMs, Gemini-Pro (Team et al., 2023) and GPT-4V(ision) (OpenAI, 2023). Experimental results reveal that most VLMs rarely hesitate to respond to the wrong option even though their accuracies for standard questions are adequate. Although GPT-4V and LLaVA-NeXT-34B in general perform better than other VLMs on UPD problems, they still have their own limitations in certain abilities and settings. As solutions for UPD, we explore a simple, *training-free* solution and *training-based* solution. These solutions have led to improvements in UPD performance, notable challenges remain, particularly in the AAD setting and with smaller VLMs. Our results underscore the complexity of the UPD challenge and emphasize the necessity for more innovative approaches in future research. (We refer readers to Appendix B for a detailed result and discussion for these solutions.)

2 Problem Definition

In this section, we introduce the concept of Unsolvable Problem Detection (UPD), a task designed to evaluate models' capacity to not blindly offer incorrect answers when presented with unsolvable problems. Considering various discrepancies among the provided image, question, and answer options, we categorize UPD into three distinct problem types: Absent Answer Detection (AAD), Incompatible Answer Set Detection (IASD), and Incompatible Visual Question Detection (IVQD). The details of each setting are as follows:

Absent Answer Detection (AAD): AAD tests the model's capability to recognize when the correct answer is absent from the provided choices. It challenges the model to not only analyze the content of questions and images but also identify when it cannot select a correct response due to the absence of an appropriate option.

Incompatible Answer Set Detection (IASD): IASD studies the model's ability to identify situations where the set of answer choices is incompatible with the context. Differing from AAD, in which the answer set is related to the question or the image, IASD deals with answer sets that are entirely irrelevant, challenging the model to withhold a response due to the lack of reasonable options.

Table 1: **Comparison results of the overall Dual accuracy** for the base setting, additional-option setting, and additional-instruction setting. "Original Standard" refers to the standard accuracy when using LLaVA's prompt (Liu et al., 2023a) specialized for performance improvement. **The "Original Standard" value is not Dual accuracy, but we consider it as the upper bound of Dual accuracy.** It is found that effective methods vary with VLMs. Also, the gaps from the Original Standard are clear.

		LLaVA 1.5	CogVLM	Qwen-VL	Gemini-Pro	LLaVA NeXT-13B	LLaVA NeXT-34B	GPT-4V
AAD	Original Standard	74.4	71.5	68.5	72.7	76.7	84.3	80.0
	Base	0.6	0.5	17.8	25.7	18.3	53.2	49.0
	Option	**38.8**	**39.3**	17.2	40.1	18.2	29.9	50.5
	Instruction	37.1	3.8	**25.7**	**44.6**	**38.8**	**55.2**	**56.1**
IASD	Original Standard	70.8	67.7	64.6	70.9	73.2	80.2	75.8
	Base	7.0	0.5	24.3	32.2	31.4	56.7	61.2
	Option	46.1	**18.3**	**28.4**	48.6	29.8	22.6	**65.6**
	Instruction	**52.1**	4.4	27.0	**53.6**	**57.8**	**61.9**	60.7
IVQD	Original Standard	68.8	62.9	58.4	69.1	71.3	80.9	75.3
	Base	0.0	0.0	21.6	18.8	29.8	53.4	**62.4**
	Option	**39.3**	**19.4**	**29.5**	57.3	37.9	50.6	61.5
	Instruction	31.7	9.0	28.9	**60.1**	**54.2**	**72.5**	57.9

Incompatible Visual Question Detection (IVQD): IVQD evaluates the VLMs' capability to discern when a question and image are irrelevant or inappropriate. This setting tests the model's understanding of the alignment between visual content and textual questions, aiming to spot instances where image-question pairs are incompatible.

3 MM-UPD: Benchmarks and Evaluations

MM-UPD Bench. We create AAD, IASD, and IVQD benchmarks based on MMBench (dev) (Liu et al., 2023c). MMBench (Liu et al., 2023c) is a systematically-designed objective benchmark for evaluating various abilities of vision-language models. We follow MMBench on the definition of each ability (*e.g.,* "Coarse Perception: Image Scene" and "Logic Reasoning: Future Prediction"). Based on MMBench, we create three benchmarks: (i) **MM-AAD Bench**: a dataset where the correct answer option for each question is removed. Our MM-AAD Bench has 820 AAD questions over 18 abilities. (ii) **MM-IASD**: a dataset where the answer set is completely incompatible with the context specified by the question and the image. Our MM-IASD Bench has 919 IASD questions over 18 abilities. (iii) **MM-IVQD**: a dataset where the image and question are incompatible. Our MM-IVQD Bench has 356 IVQD questions over 12 abilities. More detailed information for the preprocessing of each benchmark is provided in Appendix C.

Evaluation Metrics. Ideal VLMs should yield not only correct answers in the standard setting (where the image, question, and answer sets are all aligned, and the ground-truth answer is always within the options) but also be able to withhold answering in the UPD scenario where technically the question becomes unsolvable. Fig. G in Appendix D shows examples of these standard and UPD settings. To better reflect the ideal behavior of VLMs, we measure several metrics throughout the paper: (i) **Standard accuracy**: The accuracy on standard questions.[1] (ii) **UPD (AAD/IASD/IVQD) accuracy**: The accuracy of AAD/IASD/IVQD questions, i.e., the correct rate for questions in AAD/IASD/IVQD. (iii) **Dual accuracy**: The accuracy on standard-UPD pairs, where we count success only if the model is correct on both the standard and UPD questions.

Evaluation Settings. We test in three settings: (i) **Base setting.**: No instructions are provided to the model to withhold answers. (ii) **Additional-Option setting**: We add extra option "None of the above" for AAD and IASD and "The image and question are irrelevant." for IVQD, respectively. (iii) **Additional-Instruction setting**: We add additional *instruction* to explicitly gear the model towards acknowledging the unsolvable problem. The instruction is "If all the options are incorrect, answer F. None of the above." for AAD and IASD and "If the given image

[1]Please do not confuse "Standard accuracy" with "original standard". They use subtly different prompts. See the beginning of Appendix. B.1.

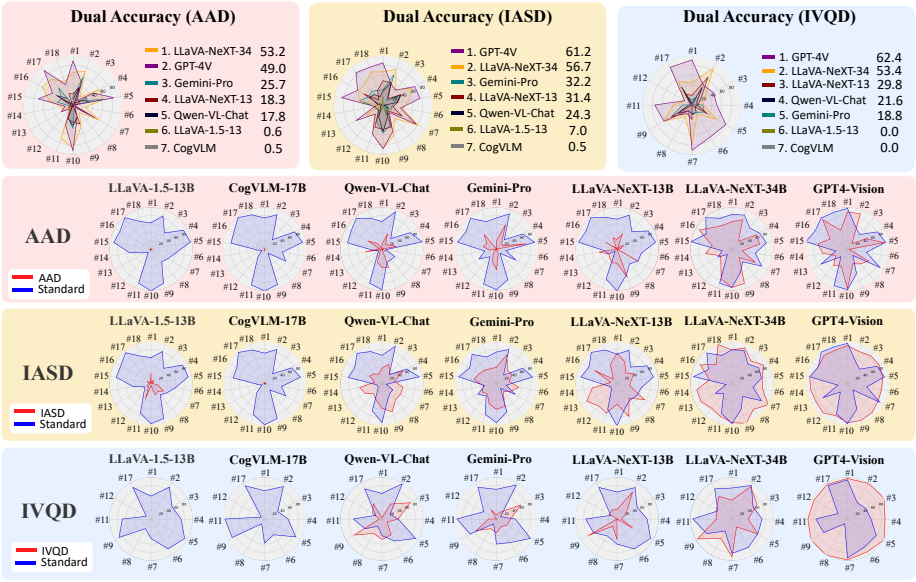


Figure 2: **Comparison results for the base setting.** Most VLMs exhibit significantly low UPD performances (red). GPT-4V and LLaVA-NeXT-34B still have their own limitations for certain abilities.

is irrelevant to the question, answer F. The image and question are irrelevant." for IVQD, respectively. The examples for each evaluation setting are shown in Fig. G in Appendix D.

4 Evaluation Results

We evaluate five state-of-the-art open-source VLMs, including LLaVA-1.5-13B (Chung et al., 2022; Liu et al., 2023a), CogVLM-17B (Wang et al., 2023d), Qwen-VL-Chat (Bai et al., 2023), the more recent LLaVA-NeXT (13B, 34B) (Liu et al., 2024c), and two close-source VLMs, Gemini-Pro (Team et al., 2023) and GPT-4V(ision) (OpenAI, 2023) (`gpt-4-vision-preview`).

Most VLMs hardly hesitate to answer. Table 1 shows the overall Dual accuracies for all settings. We find that the gaps between the Dual accuracies and the upper bound accuracies (the scores of Original Standard) are clear, which indicates the difficulty of our UPD challenge. For the base setting, we find that LLaVA-Next-34B and GPT-4V achieve higher performances than other VLMs. GPT-4V was explicitly evaluated to refuse to predict for the safety (OpenAI, 2023), and it might contribute to the higher UPD performance. LLaVA-NeXT (Liu et al., 2024c) also uses the response with GPT-4V for training, which might improve the UPD accuracy. However, there is still a performance gap from the upper bound scores.

The performance tendency differs a lot by ability. In Fig. 2, we show the radar charts based on each ability for the base setting. We find that the performance differs in each ability. For instance, GPT-4V and LLaVA-NeXT-34B still have their own limitations in certain abilities, *e.g.,* attribute comparison for GPT-4V and object localization for LLaVA-NeXT-34B. The radar charts and discussions in other scenarios (options, instructions, and instruction-tuning) are included in Appendix B.

5 Conclusions

This paper introduces a novel challenge, Unsolvable Problem Detection for VLMs. Our findings from experimental results show that most VLMs face significant challenges when tested against our benchmarks. This includes recent advanced models like GPT-4V and LLaVA-NeXT-34B, which also exhibit certain shortcomings in specific abilities. We hope that our task and findings will assist the research field in the trustworthiness of VLMs and promote future advances.

Acknowledgement. This research has been partially supported by JST JPMJCR22U4.

REFERENCES

Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. *arXiv preprint arXiv:2312.02896*, 2023.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. In *EMNLP*, 2018.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Ernest Davis. Unanswerable questions about images and texts. *Frontiers in Artificial Intelligence*, 3:51, 2020.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *ICLR*, 2022.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *AAAI*, 2022.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *CVPR*, 2024.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. In *CVPR*, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI*, 2024.

Yanyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. Unanswerable visual question answering. *arXiv preprint arXiv:2310.10942*, 2023.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*, 2022.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023a.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, 2024.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. In *ICML*, 2023b.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *CVPR*, 2024a.

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. *arXiv preprint arXiv:2403.04306*, 2024b.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *CVPR*, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3 it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023b.

Mengdi Li, Cornelius Weber, and Stefan Wermter. Neural networks for detecting irrelevant questions during visual question answering. In *ICANN*, 2020.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023c.

Zongxia Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *arXiv preprint arXiv:2307.05052*, 2023d.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 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 *ECCV*, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. In *ICLR*, 2024a.

Fuxiao Liu, Hao Tan, and Chris Tensmeyer. Documentclip: Linking figures and main body text in reflowed documents. In *ICPRAI*, 2024b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024c. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023c.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023d.

Chaochao Lu, Chen Qian, Guodong Zheng, Hongxing Fan, Hongzhi Gao, Jie Zhang, Jing Shao, Jingyi Deng, Jinlan Fu, Kexin Huang, Kunchang Li, Lijun Li, Limin Wang, Lu Sheng, Meiqi Chen, Ming Zhang, Qibing Ren, Sirui Chen, et al. From gpt-4 to gemini and beyond: Assessing the landscape of mllms on generalizability, trustworthiness and causality through four modalities. *arXiv preprint arXiv:2401.15071*, 2024a.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024b.

Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. The promise of premise: Harnessing question premises in visual question answering. In *EMNLP*, 2017.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *ACL*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022a.

Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *AAAI*, 2022b.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Can pre-trained networks detect familiar out-of-distribution data? *arXiv preprint arXiv:2310.00847*, 2023a.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *NeurIPS*, 2023b.

Sina Mohseni, Haotao Wang, Chaowei Xiao, Zhiding Yu, Zhangyang Wang, and Jay Yadawa. Taxonomy of machine learning safety: A survey and primer. *ACM Computing Surveys*, 55(8):1–38, 2022.

Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin F Yang, and Kai-Wei Chang. Metavl: Transferring in-context learning ability from language models to vision-language models. *arXiv preprint arXiv:2306.01311*, 2023.

OpenAI. Gpt-4v(ision) system card. 2023.

Dongmin Park, Zhaofang Qian, Guangxing Han, and Ser-Nam Lim. Mitigating dialogue hallucination for large multi-modal models via adversarial instruction tuning. *arXiv preprint arXiv:2403.10492*, 2024.

Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. *arXiv preprint arXiv:2402.13220*, 2024.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *ACL*, 2018.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *TACL*, 7:249–266, 2019.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022.

Xiangxi Shi and Stefan Lee. Benchmarking out-of-distribution detection in visual question answering. In *WACV*, 2024.

Elior Sulem, Jamaal Hay, and Dan Roth. Yes, no or IDK: The challenge of unanswerable yes/no questions. In *NAACL*, 2022.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023b.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023c.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS Datasets and Benchmarks Track*, 2023a.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? In *NeurIPS*, 2021.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV*, 2023b.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023c.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023d.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *TMLR*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022b.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 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.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. In *NeurIPS Datasets and Benchmarks Track*, 2022.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *IJCV*, 131(10):2607–2622, 2023a.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. *arXiv preprint arXiv:2312.07000*, 2023b.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023a.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2023b.

Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *ICCV*, 2019.

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. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023a.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*, 2024.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023b.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023c.

Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023a.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. In *ICLR*, 2024.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023b.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *NeurIPS*, 2023c.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.

APPENDIX

A RELATED WORK

Vision Language Model (VLM). Pivotal aspects of VLM research revolve around instruction tuning (Liu et al., 2023b; Zhu et al., 2024; Wang et al., 2023d) and multimodal context learning (Alayrac et al., 2022; Awadalla et al., 2023; Li et al., 2023a). For instruction tuning, with opensource instruction-tuned LLMs, such as FLAN-T5 (Chung et al., 2022), LLaMA (Touvron et al., 2023b;c) and Vicuna (vic, 2023), VLMs, such as LLaVA (Liu et al., 2023b;a; 2024c), MiniGPT-4 (Zhu et al., 2024), CogVLM (Wang et al., 2023d), utilized open-source resources and improved the instruction-following capabilities of VLMs. The development of VLMs has led to significant improvements in both the amount and quality of visual instructional data. LLaMA-Adapter (Gao et al., 2023; Zhang et al., 2024), mPlug-OWL (Ye et al., 2023), SVIT (Zhao et al., 2023a), LRV-Instruction (Liu et al., 2024a), and InstructBLIP (Dai et al., 2023) are the models for these developments. Multi-modal in-context learning has been explored in depth by models such as Flamingo (Alayrac et al., 2022) and OpenFlamingo (Awadalla et al., 2023), Otter (Li et al., 2023a), M3IT (Li et al., 2023b), MetaVL (Monajatipoor et al., 2023), Sparkles (Huang et al., 2023b), and MMICL (Zhao et al., 2024). These models have significantly contributed to the progress in multimodal training and instruction-following capabilities. In this work, we evaluate the trustworthiness of these powerful VLMs with our UPD benchmarks.

VLM Benchmarks. As multi-modal pretraining and instruction tuning have gained prominence, the previous standard evaluation benchmarks *e.g.,* VQA (Antol et al., 2015; Goyal et al., 2017), OK-VQA (Marino et al., 2019), MSCOCO (Lin et al., 2014), and GQA (Hudson & Manning, 2019) become insufficient. Consequently, various comprehensive benchmarks have been developed to evaluate the diverse capabilities of VLMs. These benchmarks encompass a broad range of VLMs' skills such as OCR (Liu et al., 2023d), adversarial robustness (Zhao et al., 2023c), image quality (Wu et al., 2024), and hallucination (Cui et al., 2023; Guan et al., 2024; Sun et al., 2023). Additionally, more extensive evaluations have been carried out through benchmarks such as LAMM (Yin et al., 2023b), LLMM-eHub (Xu et al., 2023), SEED (Li et al., 2024), LLaVA-Bench (Liu et al., 2023b) MMBench (Liu et al., 2023c), MM-Vet (Yu et al., 2023), MathVista (Lu et al., 2024b) and MMMU (Yue et al., 2024), offering a more holistic assessment of VLMs. As LLMs and VLMs are deployed across increasingly diverse domains, concerns are simultaneously growing about their trustworthiness (Zhao et al., 2023c; Cui et al., 2023; Guan et al., 2024; Sun et al., 2023; Wang et al., 2023a). For LLM, comprehensive studies have been conducted (*e.g.,* toxicity, stereotype bias, adversarial robustness, OOD generalization) (Wang et al., 2023a). For VLMs, adversarial robustness (Zhao et al., 2023c; Tu et al., 2023), OOD generalization (Tu et al., 2023), cross-style visual capability (Cai et al., 2023), and VLM hallucination (Jiang et al., 2024a; Li et al., 2023c; Gunjal et al., 2024; Liu et al., 2024a; Guan et al., 2024) have been investigated. Unlike these existing studies, we provide novel benchmarks for Unsolvable Problem Detection which examines the ability to identify unsolvable problems.

Model Hallucinations. In VLMs, "hallucination" typically refers to situations where the generated responses contain information that is inconsistent in the visual content (Rohrbach et al., 2018; Wang et al., 2023c; Zhou et al., 2024; Guan et al., 2024; Sun et al., 2023; Cui et al., 2023; Jiang et al., 2024a). Recent VLMs, such as LLaVA (Chung et al., 2022; Liu et al., 2023a), have also encountered the challenge of hallucination (Jiang et al., 2024a). To evaluate hallucination in VLMs, various benchmarks, POPE (Li et al., 2023c), M-HalDetect (Gunjal et al., 2024), GAVIE (Liu et al., 2024a), HallusionBench (Guan et al., 2024), and Bingo (Cui et al., 2023) have been proposed. Hallucination evaluation and detection (Li et al., 2023c; Wang et al., 2023c; Liu et al., 2024a), and hallucination mitigation (Yin et al., 2023a; Zhou et al., 2024; Gunjal et al., 2024; Liu et al., 2024a; Favero et al., 2024; Huang et al., 2024; Park et al., 2024; Wang et al., 2024) have also been explored. These existing studies deal with a wide range of hallucination issues. Unlike prior works, our work focuses on evaluating the VLMs' ability to hesitate to answer when faced with unsolvable problems. While some studies in the LLM (Feng et al., 2024; Kadavath et al., 2022; Yang et al., 2023b) examine the ability to make LLMs answer "I don't know" based on the existence of the corresponding parametric knowledge, UPD is a task that recognizes whether the given problem is solvable. In line with a similar motivation to our study, concurrently, Qian *et al.* (Qian et al., 2024) and Jiang *et al.* (Jiang

et al., 2024b) test performances under the incompatibility of the image and question (IVQD). The main difference from concurrent work is (i) single-choice questions (an important question format alongside the free description form (Qian et al., 2024; Jiang et al., 2024b)), (ii) datasets with more diverse questions and abilities, (iii) definition of three kinds of problems of AAD, IASD, and IVQD. In this paper, we focus on UPD, a specific issue within hallucination, and provide a comprehensive and systematic problem definition and benchmarking.

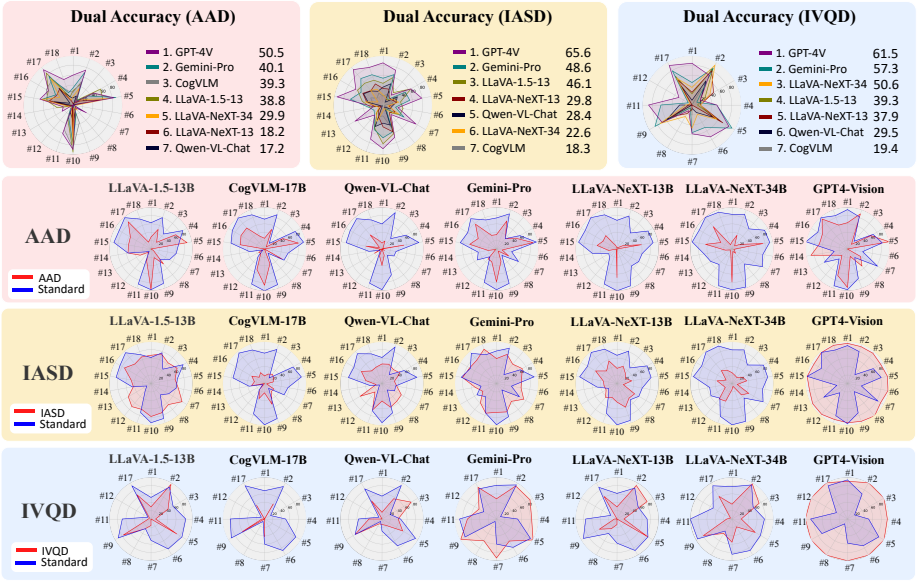
AI Safety. A reliable visual recognition system should not only produce accurate predictions on known context but also detect unknown examples (Amodei et al., 2016; Mohseni et al., 2022; Hendrycks et al., 2021; Hendrycks & Mazeika, 2022). The representative research field to address this safety aspect is out-of-distribution (OOD) detection (Hendrycks & Gimpel, 2017; Liang et al., 2018; Yang et al., 2021; 2022; Zhang et al., 2023a). OOD detection is the task of detecting unknown samples during inference to ensure the safety of the in-distribution (ID) classifiers. Along with the evolution of the close-set classifiers, the target tasks for OOD detection have evolved from the detectors for conventional single-modal classifiers to recent CLIP-based methods (Hendrycks & Gimpel, 2017; Yu & Aizawa, 2019; Wang et al., 2021; Du et al., 2022; Ming et al., 2022b; Esmaeilpour et al., 2022; Ming et al., 2022a; Yang et al., 2023a; Wang et al., 2023b; Miyai et al., 2023a;b). The next crucial challenge is to evolve the problems faced in OOD detection to VLMs in the VQA task. We consider that our UPD is an extension of the concept of OOD detection, where the model should detect and not predict unexpected input data. Unlike OOD detection with conventional task-specific VQA models (Shi & Lee, 2024), UPD targets VLMs with large amounts of knowledge. Therefore, UPD considers the discrepancies among the given image, question, and options rather than the previous notion of distribution. UPD extends OOD detection for VLMs, enabling it to handle a wider range of tasks beyond specific tasks to ensure the safety of VLMs' applications.

Unanswerable Questions for Question Answering. Unanswerable questions have been addressed in the field of natural language processing (NLP), such as single-round QA (Rajpurkar et al., 2018), multi-round dialogues (Choi et al., 2018; Reddy et al., 2019), binary questions (Sulem et al., 2022). Inspired by developments in the field of NLP, some existing studies have addressed unanswerable questions for VQA (Mahendru et al., 2017; Li et al., 2020; Davis, 2020; Guo et al., 2023). However, the main target of these studies (Mahendru et al., 2017; Li et al., 2020; Davis, 2020) is to examine the performance of traditional task-specific VQA models, so these are insufficient to examine the generic models such as the current VLMs. The most recent work (Guo et al., 2023) examines the performance of BLIP (Li et al., 2022), but addresses only the setting for image and question incompatibility (IVQD) and does not cover other problems. In addition, (Guo et al., 2023) lacks a detailed analysis based on ability and comparisons with more recent VLMs. Unlike previous work, our main focus is to evaluate modern generic VLMs' performances systematically. To achieve this, we propose three types of problems, AAD, IASD, and IVQD, and provide benchmarks on fine-grained abilities. These proposals are ensured by the fact that VLMs exhibit different tendencies for each problem setting and for each specific ability.

B Detail Results and Discussions

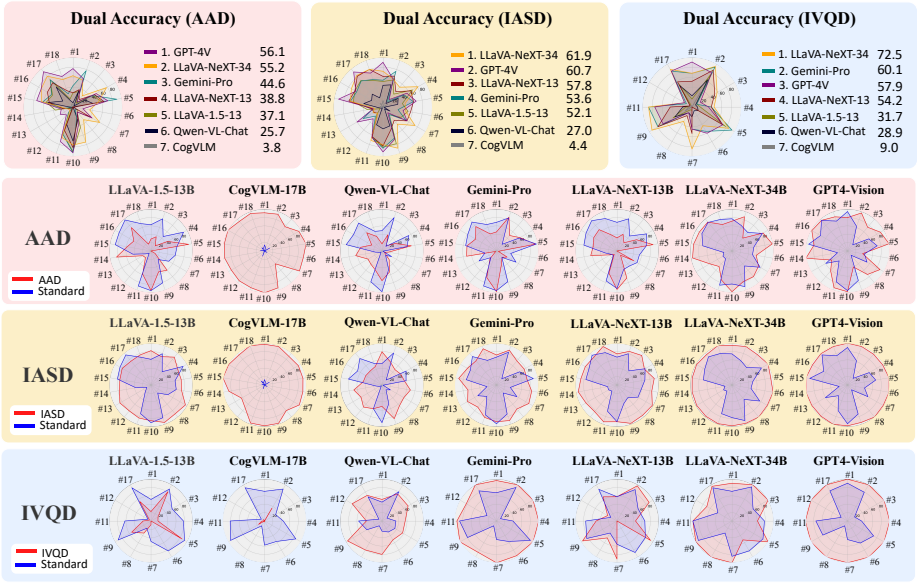
B.1 Investigation I: Training-Free Approaches

In this section, we study the effect of prompt engineering-based training-free approaches on UPD, specifically by comparing the results across the base, additional-option, and additional-instruction settings (since the latter two include extra prompts). The overall Dual accuracies for each setting are presented in Table 1 (main). To explicitly demonstrate the difficulty of UPD problems, we aim to find the "upper bound" of Dual accuracy, which shall be the maximum standard accuracy according to the definition of Dual accuracy. Therefore, we follow the LLaVA-1.5 code and use the additional prompt "Answer with the option' s letter from the given choices directly." to reach for high standard accuracy value, shown in Fig. G (d). However, it is worth noting that while such a prompt is specialized only for standard accuracy, it yields significantly worse UPD performance according to our preliminary experiments. So, in our actual UPD experiments, we explore our base, additional-option, and additional-instruction prompt with prompts in Fig. G (a-c) instead of LLaVA's original prompt.



#1: OCR #2: Celebrity Recognition #3: Object Localization #4: Attribute Recognition #5: Action Recognition #6: Attribute Comparison #7: Nature Relation #8: Physical Relation #9: Social Relation #10: Identity Reasoning #11: Function Reasoning #12: Physical Property Reasoning #13: Structuralized Image-text Understanding #14: Future Prediction #15: Image Topic #16: Image Emotion #17: Image Scene #18: Image Style

Figure A: **Comparison results for the setting with additional options.** Even though most VLMs improve the performance by adding options, the performances are still not sufficient, especially for AAD.



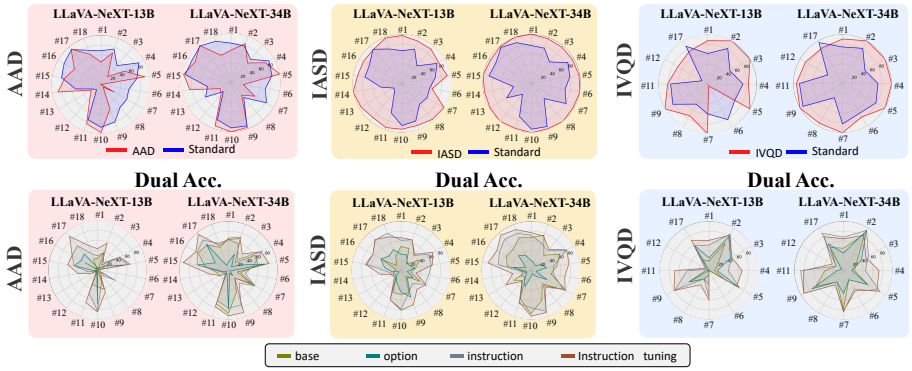
#1: OCR #2: Celebrity Recognition #3: Object Localization #4: Attribute Recognition #5: Action Recognition #6: Attribute Comparison #7: Nature Relation #8: Physical Relation #9: Social Relation #10: Identity Reasoning #11: Function Reasoning #12: Physical Property Reasoning #13: Structuralized Image-text Understanding #14: Future Prediction #15: Image Topic #16: Image Emotion #17: Image Scene #18: Image Style

Figure B: **Comparison results for the setting with an additional instruction.** Adding instruction improves the UPD performances (red), while the standard performances (blue) degrade.

B.1.1 Findings on the Base Setting

We summarize the results with radar plots for this setting in Fig. 2 (main).

Most VLMs hardly hesitate to answer for the base setting. In Fig. 2 (main), we show the results for the base setting. The crucial finding is that most VLMs, LLaVA-1.5, CogVLM, Qwen-VL-Chat, and Gemini-Pro rarely hesitate to answer and have answered from incorrect options in all AAD, IASD, and IVQD settings. In particular, MM-IASD reveals that LLaVA-1.5 and CogVLM have



#1: OCR #2: Celebrity Recognition #3: Object Localization #4: Attribute Recognition #5: Action Recognition #6: Attribute Comparison #7: Nature Relation #8: Physical Relation #9: Social Relation #10: Identity Reasoning #11: Function Reasoning #12: Physical Property Reasoning #13: Structurialized Image-text Understanding #14: Future Prediction #15: Image Topic #16: Image Emotion #17: Image Scene #18: Image Style

Figure C: **Results for instruction tuning** with LLaVA-NeXT-13B and 34B. Instruction tuning is the most effective among the comparison methods.

low inherent capacities to hesitate answers even though the answer-set is completely wrong. These VLMs have adequate standard performance for existing MMBench benchmarks (blue parts). Therefore, these results show that our benchmark reveals a new dimension for the VLMs' performances.

LLaVA-NeXT-34B and GPT-4V are still vulnerable in some abilities and settings. We find that LLaVA-NeXT-34B and GPT-4V achieve higher performances than other VLMs. GPT-4V was explicitly evaluated to refuse to predict for the safety (OpenAI, 2023), and we consider it contributes to the higher UPD performance. LLaVA-NeXT (Liu et al., 2024c) also uses the response with GPT-4V for training, which might improve the UPD accuracy. However, they still have some limits to their performance. For the AAD setting, GPT-4V has lower AAD performances in #4 Attribute Recognition, #6 Attribute Comparison, #9 Social Relation, #11 Function Reasoning, #16 Image Emotion even though the standard performances in their abilities are adequate. LLaVA-NeXT-34B also struggles in the abilities of #3 Object Localization, #6 Attribute Comparison, and #15 Image Topic. For the IVQD setting, we find that LLaVA-NeXT-34B and GPT-4V show different trends, and LLaVA-NeXT-34B has worse ability to refrain from answering than GPT-4V.

B.1.2 Findings on the Additional-Option Setting

We summarize the results with radar plots for this setting in Fig. A.

Adding option is effective for LLaVA-1.5 and CogVLM. Table 1 shows that adding an option is more effective for LLaVA-1.5 and CogVLM than adding instructions. This provides an interesting finding that effective prompt engineering methods for UPD differ for models.

LLaVA-NeXTs perform badly with additional options. As shown in Fig. A, LLaVA-NeXTs degrade the UPD performance a lot by adding additional option. For LLaVA-NeXT-13B and 34B, they do not choose the option for unsolvable questions ("None of the above" for AAD and IASD). The detailed training recipe for LLaVA-NeXT is still unpublic, but the reason for this low score might be that the choice-type questions in the LLaVA-NeXT's training data do not include UPD-like data where "None of the above" is the answer, so they are very unlikely to choose that option despite that it is the correct one for UPD questions.

Performances are still not sufficient. Even though the performances of some VLMs can be improved by adding additional options, the AAD, IASD, and IVQD accuracy are still lower than standard accuracy. Even for GPT-4V, the performances for #6 Attribute Comparison, #7 Nature Relation, #9 Social Relation, #11 Function Reasoning are still low in AAD.

B.1.3 Findings on the Additional-Instruction Setting

We summarize the results with radar plots for this setting in Fig. B.

Table A: Comparison results with the overall Dual accuracy for instruction tuning.

(a) LLaVA-NeXT-13B

	Base	Opt	Inst	Inst Tuning
AAD	18.3	18.2	38.8	**47.6**
IASD	31.4	29.8	57.8	**60.0**
IVQD	29.8	37.9	54.2	**59.6**

(b) LLaVA-NeXT-34B

	Base	Opt	Inst	Inst Tuning
AAD	53.2	29.9	55.2	**63.8**
IASD	56.7	22.6	61.9	**73.3**
IVQD	53.4	50.6	**72.5**	70.2

Adding instruction is effective for Gemini-Pro and LLaVA-NeXTs. Table. 1 shows that the overall Dual accuracies with additional instruction are higher for Gemini-Pro and LLaVA-NeXTs. As for other VLMs, by adding instructions, the standard accuracies (blue) become much lower, which degrades the dual accuracies.

Additional instruction improves the UPD (red) accuracy. Compared to the setting with additional options, the UPD accuracies (red) are higher when we add additional instructions. In particular, LLaVA-NeXTs achieve higher UPD accuracies, even though they do not perform well when given options. When looking into IASD, we find that most VLMs can correctly answer IASD questions. For AAD and IVQD, large-scale LLaVA-NeXT-34B and GPT-4V have adequate UPD performances (red). On the other hand, other VLMs, except for CogVLM, still have some difficulties.

Additional instruction degrades the standard (blue) accuracy. Although additional instruction improves the UPD accuracy, it degrades the standard (blue) accuracy for most VLMs. This is due to the fact that VLMs regard even standard questions as unsolvable ones. CogVLM is particularly remarkable for this phenomenon and it answers "None of the above" even for most standard questions in the AAD and IASD scenarios. As for other VLMs, the performances of standard questions are lower than those in other settings. This illustrates the difficulty of accurately distinguishing between standard and unsolvable questions and providing appropriate answers to each.

B.2 Investigation II: Training-based Approaches

B.2.1 Experiment Setup

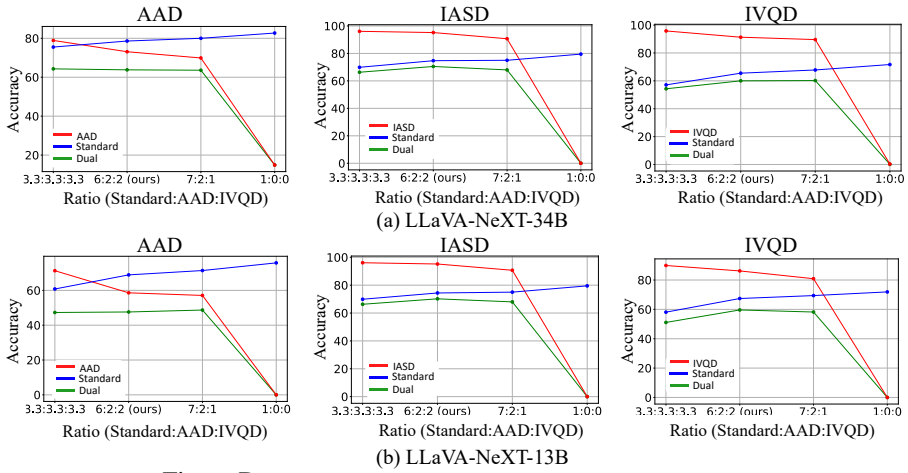
Original datasets. For the dataset, we use a subset of an open-knowledge VQA dataset, A-OKVQA (Schwenk et al., 2022). It is a single-choice type VQA dataset that has been used for training InstructBLIP (Dai et al., 2023) and LLaVA-1.5 (Liu et al., 2023a). The samples in A-OKVQA do not overlap with our benchmarks. Following LLaVA-1.5's recipe (Liu et al., 2023a), we use a specific response formatting: "Answer with the option' s letter from the given choices directly". Also, we augment each question k times, where k is the number of choices per question, to counterbalance the lack of single-choice data following LLaVA-1.5's recipe.

Instruction tuning datasets for UPD. To address all three types of problems, the ratio of the tuning data for each task is important. Therefore, we examine the difficulty and heterogeneity of each task and then seek the optimal amount and proportion of each type of question. We first create 4 kinds of datasets for standard questions, AAD questions, IASD questions, and IVQD questions, respectively. For each dataset, we include the questions for the base setting and the questions with additional options. For AAD/IASD/IVQD datasets, we set "I cannot answer." as an answer for the questions in the base setting. Also, to make it robust for the number of options, we create the questions with 2-4 options by augmentations. Through our experiments, we find that the most effective recipe is that we include 20% of AAD and IVQD questions respectively and not include IASD samples. Also, we find that 10,000 samples are enough for tuning. The reason for not including IASD data is explained in Sec. B.3, and the ablation study on ratio and data size is shown in the supplementary.

Tuning method. As for the tuning method, we adopt LoRA tuning (Hu et al., 2022) by considering the effectiveness and low memory usage.

B.2.2 Experimental Results and Findings

In Fig. C and Table A, we show the results for instruction tuning.



(a) LLaVA-NeXT-34B

(b) LLaVA-NeXT-13B

Figure D: Ablation on the ratio of Standard, AAD, and IVQD.

Table B: Difficulty and heterogeneity of each task. We use LLaVA-NeXT-34B. We find that AAD and IVQD require their own training data, while IASD can be addressed by the AAD and IVQD training data.

(a) Dual Accuracy

Training Data	AAD	IASD	IVQD
Standard+AAD	**66.5**	72.9	51.7
Standard+IASD	45.2	74.4	26.7
Standard+IVQD	52.1	72.2	**73.6**

(b) UPD Accuracy

Training Data	AAD	IASD	IVQD
Standard+AAD	**73.9**	96.4	63.8
Standard+IASD	46.7	96.1	32.0
Standard+IVQD	55.8	94.7	**95.8**

Instruction tuning is more effective than other methods. Table A shows that instruction tuning is more effective than prompt engineering for most settings. Only for IVQD with LLaVA-NeXT-34B, the performance of instruction tuning is slightly inferior to that of additional instruction, but Fig. C shows adequately high performance for IVQD.

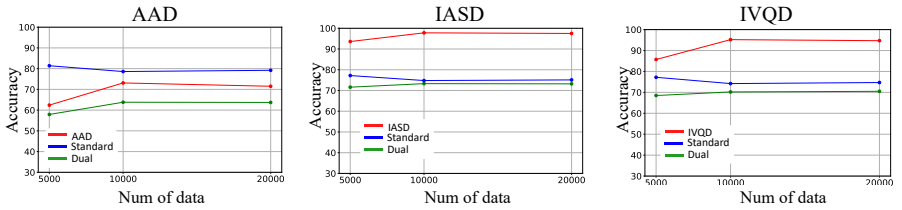
AAD is still the most challenging. Fig. C and Table A show that AAD is the most difficult in all three UPD scenarios. For example, LLaVA-NeXT-34B has large gap between standard and AAD accuracies for #13 Structuralized Image-text Understanding. This category includes questions on programming code, and we find VLMs can hardly hesitate to answer for programming code. This shows the difficulty of improving the AAD performances in certain abilities.

Smaller VLMs still have some difficulties. LLaVA-NeXT-13B also shows an improvement with instruction tuning, but the performance in some abilities still suffers. This indicates that the model size and capacity are important for UPD, and therefore, the same instruction tuning recipe does not yield as good results as larger models like LLaVA-NeXT-34B. Improving the UPD accuracies on smaller models will be one of the future challenges.

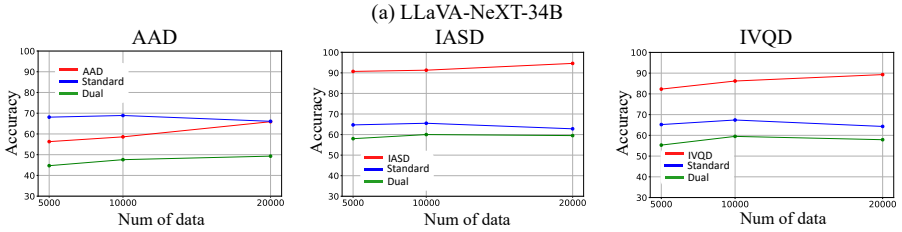
B.2.3 Ablation Study

Ablation on ratio of each UPD data. In Fig. D, we illustrate the relationship between the ratio of Standard, AAD, and IVQD instruction tuning data and the performance of each UPD, Standard, and Dual accuracy. We set the ratio of Standard: AAD: IVQD to 3.3:3.3:3.3, 6:2:2, 7:2:1, 1:0:0. From this result, increasing the ratio of UPD tuning data, the UPD performance improved much while the standard accuracy degrades. Conversely, increasing the proportion of Standard data degrades the UPD performance. We can see that the ratio of 6:2:2 is an effective ratio for all the settings.

Ablation on data size. In Fig. E, we illustrate the relationship between the tuning data size and the performance of each UPD, Standard, and Dual accuracy. In this experiment, we set the ratio of Standard, AAD, and IVQD is 0.6, 0.2, and 0.2. From this result, 10,000 samples are enough to tune for our LoRA-based instruction tuning.

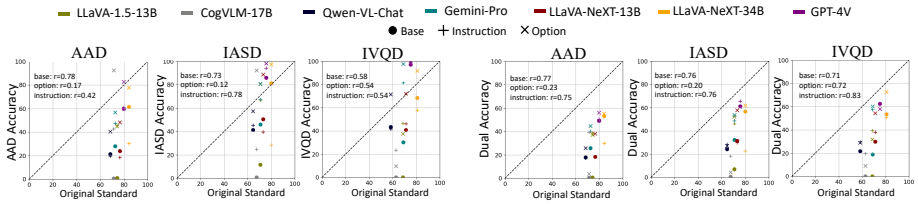


(a) LLaVA-NeXT-34B



(b) LLaVA-NeXT-13B

Figure E: Ablation on the number of instruction tuning data.



(a) Original Standard and UPD accuracy (b) Original Standard and Dual accuracy

Figure F: Correlation between Standard and UPD/Dual accuracies.

B.3 Further Analysis

Does UPD accuracy correlate with standard accuracy? Many studies on VLMs (Bai et al., 2023; Liu et al., 2023b;a; 2024c) aim to increase the Standard accuracy. To investigate whether these efforts also contribute to the improvements of UPD, we plot the performance with the upper of Standard Accuracy (Original Standard) and UPD/Dual Accuracy in Fig. F and calculate the correlation between them. The results show that while there is high correlation in the base and instruction setting, the correlation can be weak for the option setting. Meanwhile, we can see that each model has its own strengths and weaknesses, so we need to examine the performance of each model individually, which indicates the importance and difficulty of our UPD benchmarks.

Can one UPD dataset help others? To create a dataset that addresses all UPD problems, it is crucial to examine the difficulty and heterogeneity of each task. To this end, we compare the performances when we use only one UPD dataset from all three kinds of UPD datasets, which indicates the difficulty or similarity of each task. In Table B, we show the result. From this result, we find that, for AAD and IVQD, we need to include their own training data, while both IVQD and AAD data are sufficient to solve IASD questions. This is because IASD can be considered a simpler version of the AAD question since the answer-set does not include the correct answer, and it is also related to IVQD since the answer-set is not related to the given image. Hence, to reduce the complexity, we can create the tuning dataset from AAD and IVQD data without IASD data.

C Benchmark Construction

We carefully adapt MMBench to create our MM-UPD Bench. MMBench is a VLM benchmark consisting of single-choice questions covering different ability dimensions. For simplicity of explanation, we show the mapping table of each index and ability. To create the MM-UPD Bench from MMBench, we perform the following preprocessing.

C.1 Preprocess for MMBench

Before creating each MM-UPD Bench, we performed the following pre-processings for the original MMBench to ensure the quality of our benchmarks.

Exclusion of some image-agnostic questions. In the original MMBench, a subset of the questions were image-agnostic questions, which can be answered with only text information. To ensure the validity of the VLM benchmark, we carefully excluded these questions. First, we removed the questions that could be accurately answered by the GPT-4 using only text information. Then, we manually checked and excluded the remaining image-agnostic questions. In total, we removed 13% of the original questions as image-agnostic questions. Therefore, we argue that our benchmark consists of image-dependent questions.

Exclusion of Image Quality ability. In the original MMBench, the Image Quality ability questions consist of 31 two-choice questions and 22 four-choice questions. We removed the 2-choice questions in the AAD settings so that more than two choices remain after masking the choices. As for the remaining four-choice questions, our preliminary experiments indicated that these questions proved to be extremely difficult even with the original standard settings. Since it is difficult to measure accurate UPD performances with the ability that is extremely difficult even for the Standard setting, we removed the Image Quality ability.

Exclusion of options related "None of the above". We remove the questions that originally had options related "None of the above" in order to guarantee that no correct option exists after masking the correct option. Specifically, a few questions have the option of "None of these options are correct." or "All above are not right". Since these options are not correct answers for the original questions, we simply deleted such options.

Clarification of the meaning of the options. We clarify the meaning of the options. Specifically, some questions in Attribute Comparison have "Can't judge". "Can't judge" means that "I can' t judge from the image since the image does not have enough information". However, "Can't judge" might be interpreted as "Since the given options are incorrect, can't judge." Therefore, we changed the option of "Can't judge" to "Can't judge from the image due to the lack of image information" to reduce the ambiguity.

C.2 Construction of MM-AAD Bench

When creating the MM-AAD Bench, we mask the correct options and remove all questions that originally have two options (which after removal would have only one option left). Also, we remove the questions whose answer is 'both A,B, and C" and 'all of these options are correct". To ensure no answer is present in the options, we also manually remove some questions with ambiguity where one of the remaining options is very similar to the masked correct option (*e.g.,* Q. What can be the relationship of these people in this image? Masked Option: Friends, Similar remaining option: Colleagues). Our MM-AAD Bench has 820 AAD questions over 18 abilities. The distribution of questions for each ability is shown at the top of Table D.

C.3 Construction of MM-IASD Bench

To create MM-IASD, we shuffle all questions and answer sets and pair each question with a random answer set. To further ensure the incompatibility, after the shuffling, we manually removed questions where the shuffled answer set was somehow compatible with the question (*e.g.,* Q. Which of the following captions best describes this image? Correct answer: A person holding a bouquet of flowers, Similar shuffled option: Happiness). Our MM-IASD Bench has 919 IASD questions over 18 abilities. The distribution of questions for each ability is shown in the middle of Table D.

C.4 Construction of MM-IVQD Bench

To create MM-IVQD Bench, we first exclude the questions that can be relevant to most images and then shuffle the original image-question pairs. In Table E, we show some representative examples of removed questions. For example, the question of "How many ..." can be compatible with any image, since the correct option of "None of the above" always exists for any image even when the image has no corresponding objects. For the question of "What' s the profession ...", we can interpret the profession from any kind of image (*e.g.,* A beautifully captured image would suggest the profession of a photographer). In addition, we exclude the option "Can't judge from the image

Table C: Mapping table of indices and abilities

#1	#2	#3	#4	#5	#6	#7
OCR	Celebrity Recognition	Object Localization	Attribute Recognition	Action Recognition	Attribute Comparison	Nature Relation

#8	#9	#10	#11	#12	#13
Physical Relation	Social Relation	Identity Reasoning	Function Reasoning	Physical Property Reasoning	Structuralized Image-text Understanding

#14	#15	#16	#17	#18
Future Prediction	Image Topic	Image Emotion	Image Scene	Image Style

Table D: Distribution of questions per each ability.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	total
AAD	35	94	62	50	49	44	45	15	32	38	46	29	44	25	31	42	93	46	820
IASD	39	97	77	54	53	39	43	20	42	41	63	42	43	35	33	49	98	51	919
IVQD	31	68	36	18	14	23	45	15	43	-	16	23	-	-	-	-	24	-	356

Table E: Representative samples for removed questions

Ability	**Example of removed question**
#3 Object Localization	How many dogs are in this picture?
#15 Image Topic	Which one is the correct caption of this image?
#16 Image Emotion	Which mood does this image convey?
#13 Structuralized Image-text Understanding	Which Python code can generate the content of the image?
#14 Future Prediction	What will happen next?
#10 Identity Reasoning	What's the profession of the people in this picture?
#18 Image Style	Which style is represented in this image?

due to the lack of image information." because this option can be a correct answer for IVQD questions. Again, we conduct a manual check to guarantee the incompatibility of image-question pairs. Our MM-IVQD Bench has 356 IVQD questions over 12 abilities. The distribution of questions for each ability is shown in the bottom of Table D. Here, the lack of some ability (*e.g.,*#16 Image Emotion) indicates that there are many removed questions that can be applied to any image. Note that the small number of IVQD questions compared to AAD and IASD is due to our careful annotation and that even this number of questions is sufficient to show the performance difference between each VLM and method from our main experimental results.

D Evaluation

D.1 Evaluation Examples

In Fig. G, we show the examples of these standard and UPD settings. Here, for AAD, the standard scenario refers to the correct answer included in the provided answer set. For IASD, the standard scenario refers to the correct answer included in the provided answer set and the rest options are also relevant. For IVQD, given the same question and answer set, the standard scenario has a compatible image.

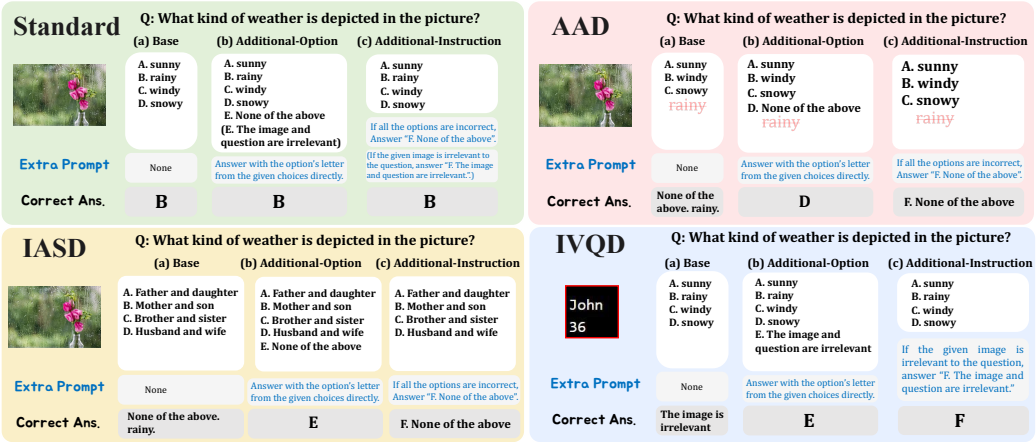


Figure G: **Examples of standard and UPD questions in each setting.**

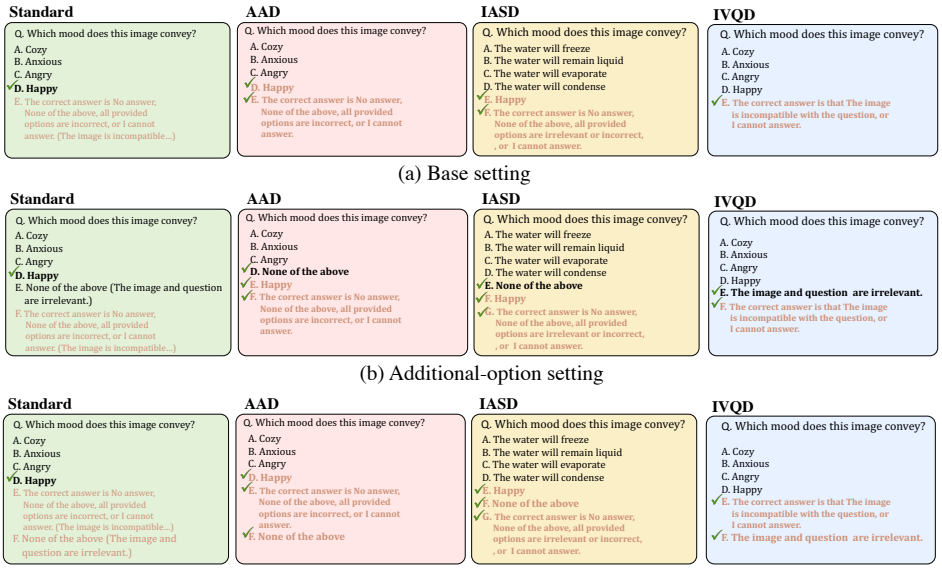


Figure H: Question and options for Chat-GPT evaluation. Brown options are additionally given to recognize UPD predictions.

D.2 Automatic Evaluation Strategy

We adopt Circular Evaluation and ChatGPT-involved Choice Extraction in MMBench (Liu et al., 2023c) as an evaluation strategy. In Circular Evaluation, a problem is tested multiple times with circularly shifted choices, and the VLM needs to succeed in all testing passes. ChatGPT-involved Choice Extraction first performs the matching algorithm and then uses ChatGPT for those that do not match.

Simplification of the matching algorithm. To apply the matching algorithm for UPD, we simplify the matching algorithm to prevent the UPD predictions from matching the given options. In detail, when an option is denoted simply by a letter such as 'A' or expressed as 'A) XXX', 'A. XXX', 'A, XXX', '(A) XXX' without the inclusion of other choices within the 'XXX' portion, it is considered that 'A' is being predicted.

Change of the template for for ChatGPT evaluation. Next, to identify the UPD prediction, we leverage ChatGPT (`gpt-3.5-turbo-0613`) following MMBench. We query it with the template, including the question, options and the corresponding VLM's prediction. We slightly change the template for the original MMBench, and create the query template for each setting in Fig. I. As for options, we add UPD-specific options to recognize UPD predictions. In Fig. H, we illustrate the

Prompt for AAD and IASD

You are an AI assistant who will help me to match an answer with several options of a single-choice question. You are provided with a question, several options, and an answer, and you need to find which option is most similar to the answer. If the meaning of all options are significantly different from the answer, output H. Your should output a single uppercase character in A, B, C, D, E, F, G (if they are valid options), and H.
Example 1: Question: What is the main object in image?
Options: A. teddy bear B. rabbit C. cat D. dog **E. no answer**
Answer: a cute teddy bear
Your output: A
Example 2: Question: What is the main object in image?
Options: A. teddy bear B. rabbit C. cat D. dog **E. None of the above**
Answer: no answer
Your output: E
Example 3:
Question: What is the main object in image?
Options: A. teddy bear B. rabbit C. cat D. dog **E. None of the above**
Answer: fish
Your output: H
Example 4:
Question: {}?
Options: {}
Answer: {}
Your output:

Prompt for IVQD

You are an AI assistant who will help me to match an answer with several options of a single-choice question. You are provided with a question, several options, and an answer, and you need to find which option is most similar to the answer. If the meaning of all options are significantly different from the answer, output H. Your should output a single uppercase character in A, B, C, D, E, F, G (if they are valid options), and H.
Example 1: Question: What is the main object in image?
Options: A. teddy bear B. rabbit C. cat D. dog **E. The image and question are irrelevant**
Answer: a cute teddy bear
Your output: A
Example 2: Question: What is the main object in image?
Options: A. teddy bear B. rabbit C. cat D. dog **E. The image and question are irrelevant**
Answer: The uploaded image and question are incompatible.
Your output: E
Example 3:
Question: What is the main object in image?
Options: A. teddy bear B. rabbit C. cat D. dog **E. The image and question are irrelevant**
Answer: fish
Your output: H
Example 4:
Question: {}?
Options: {}
Answer: {}
Your output:

Figure I: Chat-GPT query template for each setting.

options for each setting. For AAD, we add two options: a masked correct option, and the option of "The correct answer is No answer, None of the above, all provided options are incorrect, or I cannot answer.". For IASD, we add two options: a masked correct option, and the option of "The correct answer is No answer, None of the above, all provided options are irrelevant or incorrect, or I cannot answer.". For IVQD, we add an option of "The correct answer is that The image is incompatible with the question, or I cannot answer." For the additional-instruction setting, we also add the option "F. None of the above" or "F. The image and question are irrelevant.". In each setting, we regard the options indicated by check marks (Fig. H), as correct ones.

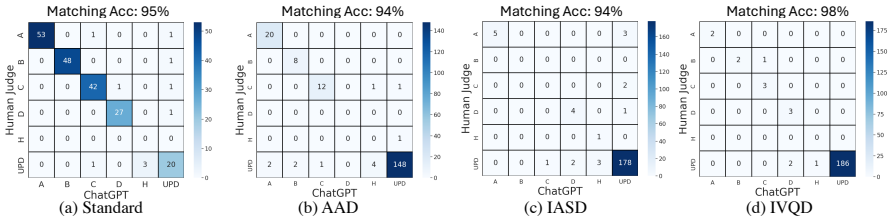
D.3 Comparison to human decision

In Fig. J, we investigate the alignment of scores given by ChatGPT and human. To investigate the performance of the UPD predictions, we sampled every 100 predictions of LLaVA-NeXT-34B and GPT-4V that were not matched by pattern matching and manually evaluated them. We found that the match rate with human evaluations is sufficiently high.

E Experimental details

E.1 Inference of VLMs

LLaVA-1.5-13B, LLaVA-NeXT-13B, LLaVA-NeXT-13B. The authors published the inference code for MMBench. Therefore, we utilize this code for our implementations. Following this code, we use a greedy decoding strategy for LLM's inference.



(a) Standard (b) AAD (c) IASD (d) IVQD

Figure J: We manually annotate the correctness of VLM' s predictions and compare its alignment with Chat-GPT

CogVLM-17B. We utilize the hugging face model (Wolf et al., 2019) for the implementations. Following this code, we use a greedy decoding strategy for LLM's inference.

Qwen-VL-Chat. We utilize the hugging face model (Wolf et al., 2019) for the implementations. In this code, they use a nucleus sampling for LLM's inference and set top-k to 0.3. Therefore, we follow this setting. Also, we tried a greedy decoding strategy, but the performance did not improve.

Gemini Pro Vision. We utilize the API of `gemini-pro-vision`. We set the temperature to 0.0 and performed a greedy decoding approach. In addition, this API filters the output based on the harm category. However, we find that such filtered outputs are irrelevant to our UPD settings. Therefore, we lower the level of safety filtering for our implementations to improve performance.

GPT-4Vision. We utilize the OpenAPI's API of `gpt-4-vision-preview` for our implementations. We set the temperature to 0.0 and performed a greedy decoding approach.

E.2 Automatic Evaluation

Following the codebase of MMBench (OpenCampass (Contributors, 2023)), we utilize Chat-GPT API (`gpt-3.5-turbo-0613`) with a temperature of 0.7 for evaluations.

F Limitation and Future Work

Exploring other methods for UPD. Another promising approach for UPD is a chain-of-thought (CoT) reasoning (Wei et al., 2022b; Kojima et al., 2022; Zhang et al., 2023c). Zero-Shot-CoT (Kojima et al., 2022) and Few-Shot-CoT (Wei et al., 2022b) can be an effective approach to LLM, but these simple methods rely on scaling LLM on the scale of 100 billion parameters (Wei et al., 2022a), making it difficult to apply them directly to the multimodal field (Zhang et al., 2023c). Therefore, it is important future work to build CoT methods for UPD.

Extension to expert-level questions. The proposed MM-UPD Bench consists of general QA datasets. However, UPD can potentially incorporate domain-specific knowledge for advanced perception and reasoning, as seen in MathVista (Lu et al., 2024b) and MMMU (Yue et al., 2024). Expanding UPD to include expert-level questions represents a significant direction for future work.

Development of post-hoc detection methods. Another direction for solving UPD is to propose post-hoc (training-free) detection methods for UPD. Proposing model-agnostic post-hoc methods can enhance the reliability of more VLMs. We consider that this post-hoc method includes not only text responses for refusals like this study but also detections with thresholding a certain score function, analogous to out-of-distribution (OOD) detection (Hendrycks & Gimpel, 2017; Liang et al., 2018; Yang et al., 2021; 2022; Zhang et al., 2023a). To propose post-hoc methods is one of the crucial future directions.

G Full Results for Each Setting

We show the full results for each setting in Table F, G, H, I, J, K, L, M, N, O, P, and Q. We provide these results via spread sheet for followers to create radar charts easily.

H FAILURE EXAMPLES

We show some failure examples for AAD, IASD and IVQD in Fig. K, L, M, N, O, and P.

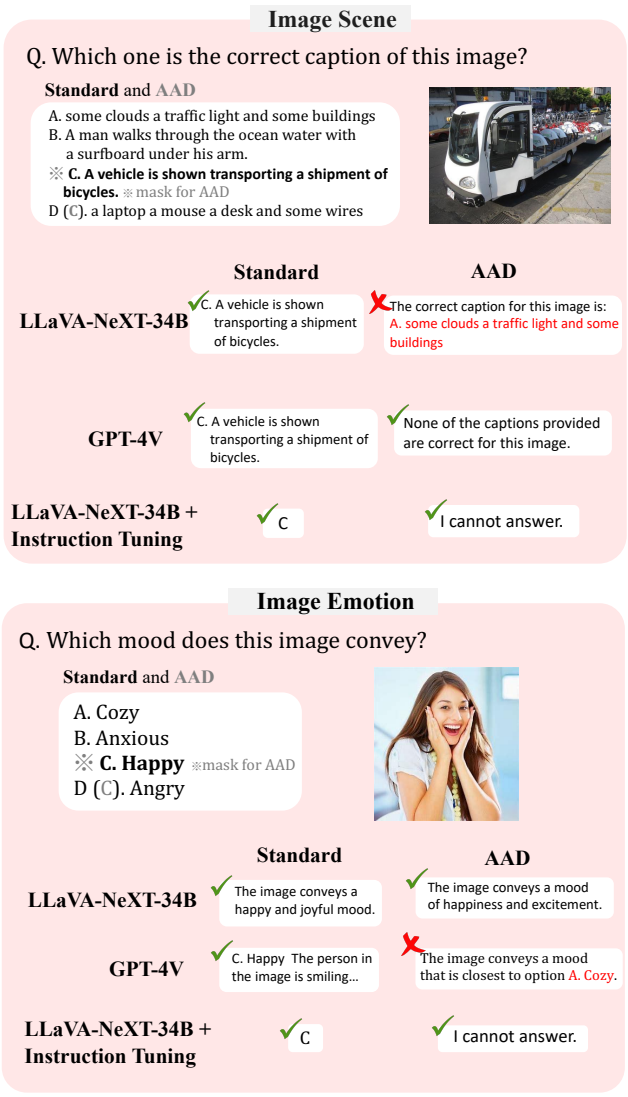


Figure K: Failure examples for AAD.

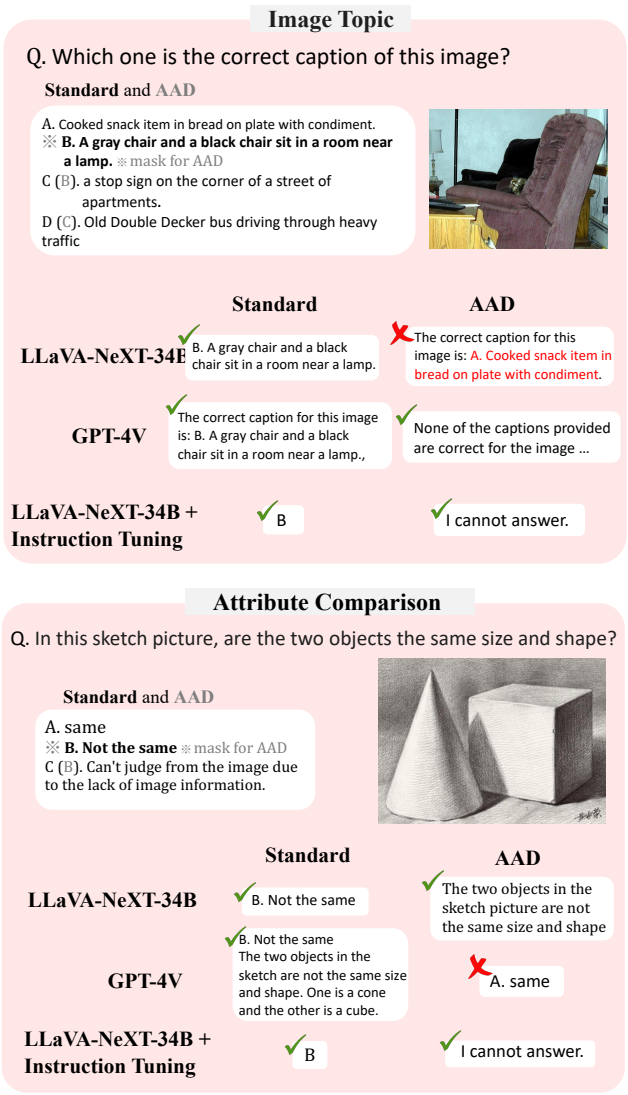


Figure L: Failure examples for AAD.

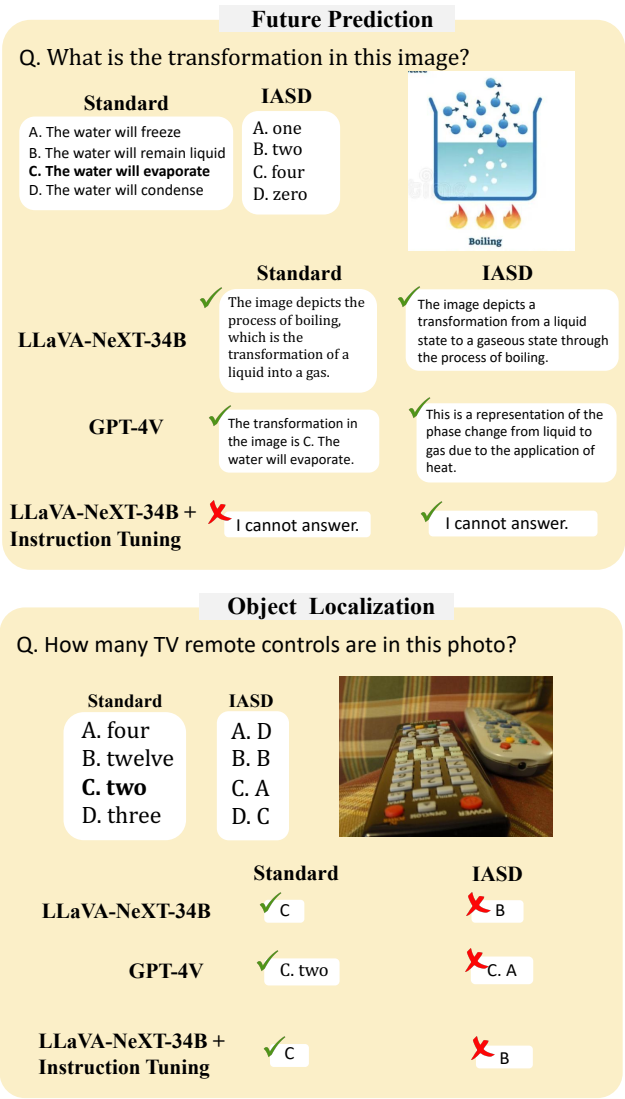


Figure M: Failure examples for IASD.

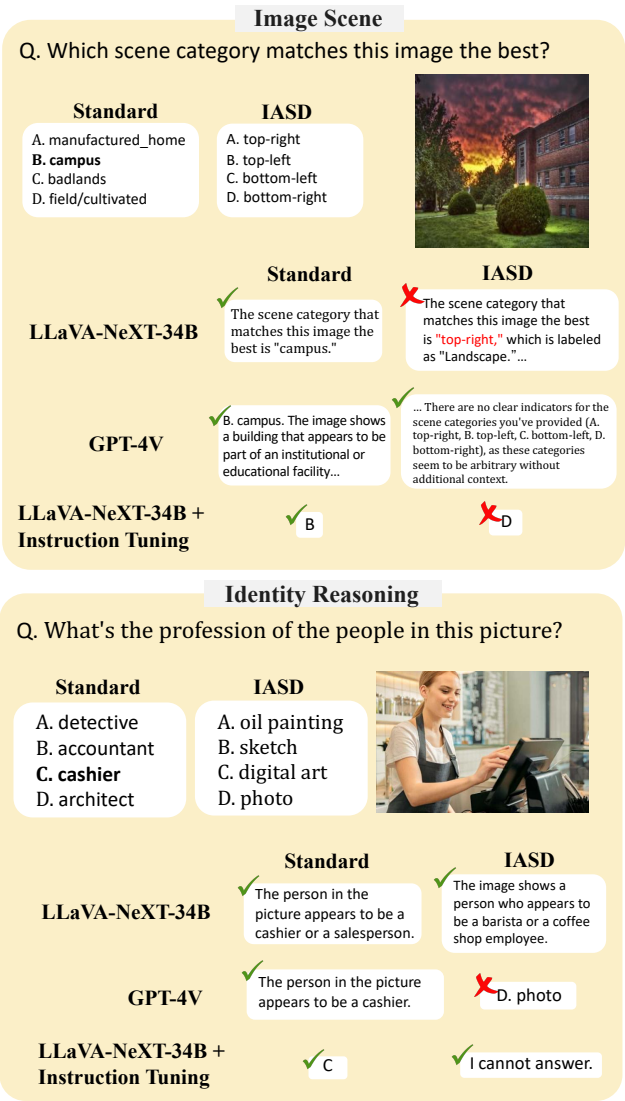
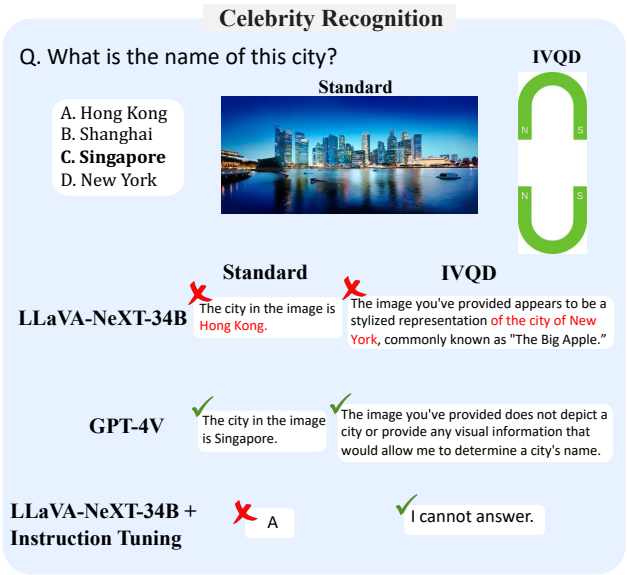


Figure N: Failure examples for IASD.



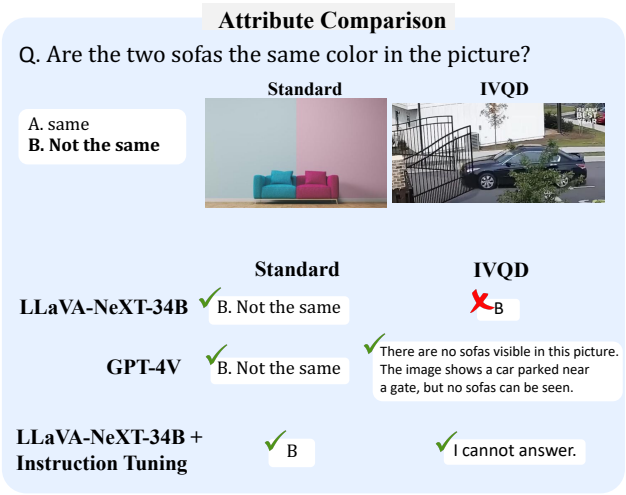


Figure O: Failure examples for IVQD.

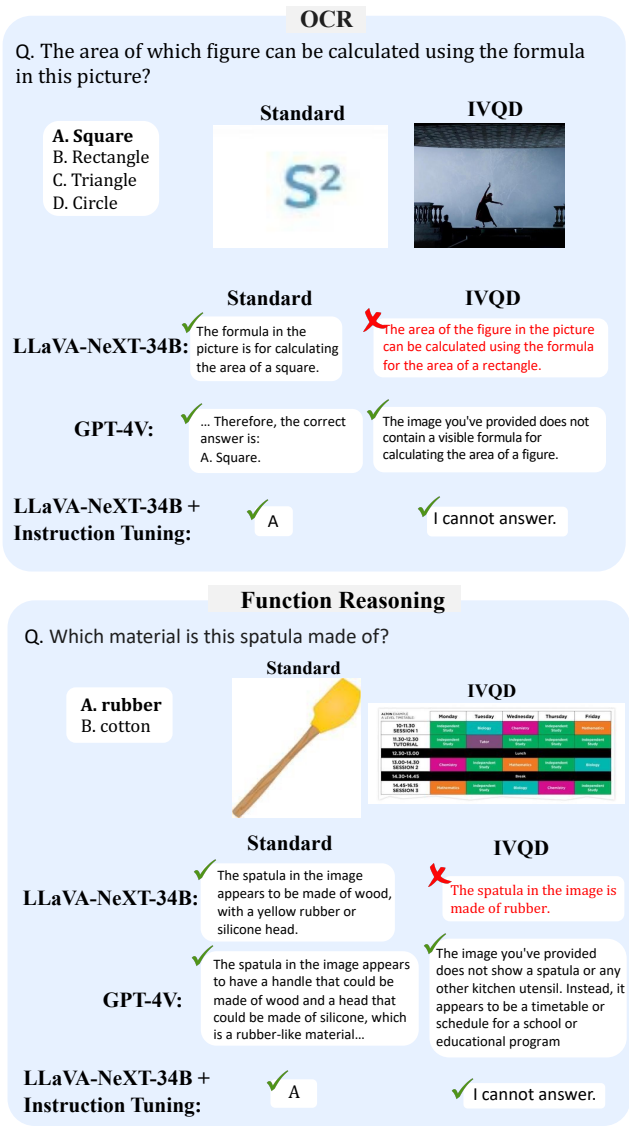


Figure P: Failure examples for IVQD.

Table F: Full results for AAD in the base setting. We report Standard accuracy, AAD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
Standard Acc.																		
LLaVA-1.5	62.9	84.0	50.0	90.0	95.9	59.1	46.7	40.0	87.5	100.0	87.0	10.3	18.2	52.0	90.3	85.7	97.8	69.6
CogVLM	74.3	87.2	29.0	82.0	91.8	43.2	57.8	13.3	90.6	100.0	91.3	24.1	6.8	16.0	80.6	83.3	97.8	87.0
Qwen-VL-Chat	65.7	94.7	32.3	78.0	98.0	9.1	35.6	33.3	34.4	97.4	87.0	17.2	9.1	20.0	83.9	85.7	95.7	78.3
Gemini-Pro	74.3	89.4	29.0	72.0	91.8	22.7	53.3	13.3	78.1	97.4	89.1	24.1	47.7	28.0	93.5	83.3	95.7	69.6
LLaVA-NeXT-13B	71.4	86.2	54.8	86.0	93.9	59.1	57.8	46.7	65.6	100.0	76.1	6.9	25.0	28.0	93.5	85.7	97.8	80.4
LLaVA-NeXT-34B	82.9	86.2	66.1	76.0	83.7	56.8	82.2	40.0	59.4	92.1	87.0	17.2	38.6	24.0	100.0	71.4	97.8	78.3
GPT-4V	97.1	56.4	45.2	82.0	91.8	45.5	88.9	20.0	62.5	97.4	93.5	24.1	72.7	28.0	100.0	88.1	96.8	93.5
AAD Acc.																		
LLaVA-1.5	0.0	0.0	0.0	0.0	2.0	0.0	2.2	0.0	0.0	0.0	4.3	3.4	2.3	4.0	0.0	0.0	0.0	0.0
CogVLM	0.0	0.0	1.6	2.0	0.0	0.0	0.0	0.0	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen-VL-Chat	11.4	38.3	12.9	20.0	34.7	0.0	11.1	0.0	46.9	44.7	32.6	6.9	2.3	40.0	6.5	38.1	17.2	0.0
Gemini-Pro	20.0	61.7	19.4	16.0	71.4	0.0	4.4	0.0	12.5	50.0	30.4	10.3	0.0	16.0	32.3	26.2	45.2	0.0
LLaVA-NeXT-13B	22.9	38.3	3.2	18.0	10.2	0.0	57.8	0.0	71.9	34.2	34.8	27.6	22.7	44.0	0.0	31.0	2.2	28.3
LLaVA-NeXT-34B	65.7	74.5	24.2	68.0	67.3	25.0	64.4	26.7	87.5	92.1	76.1	41.4	31.8	64.0	48.4	90.5	67.7	60.9
GPT-4V	88.6	89.4	16.1	50.0	87.8	4.5	57.8	46.7	37.5	94.7	43.5	51.7	31.8	60.0	71.0	47.6	91.4	47.8
Dual Acc.																		
LLaVA-1.5	0.0	0.0	0.0	0.0	2.0	0.0	2.2	0.0	0.0	0.0	4.3	0.0	0.0	4.0	0.0	0.0	0.0	0.0
CogVLM	0.0	0.0	1.6	2.0	0.0	0.0	0.0	0.0	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen-VL-Chat	11.4	38.3	4.8	20.0	34.7	0.0	8.9	0.0	25.0	42.1	28.3	0.0	0.0	4.0	6.5	38.1	17.2	0.0
Gemini-Pro	17.1	59.6	12.9	16.0	67.3	0.0	2.2	0.0	6.2	47.4	28.3	3.4	0.0	8.0	32.3	26.2	45.2	0.0
LLaVA-NeXT-13B	22.9	37.2	3.2	18.0	10.2	0.0	42.2	0.0	40.6	34.2	30.4	3.4	11.4	4.0	0.0	28.6	2.2	23.9
LLaVA-NeXT-34B	65.7	72.3	21.0	62.0	61.2	25.0	60.0	20.0	46.9	86.8	71.7	6.9	22.7	24.0	48.4	71.4	67.7	50.0
GPT-4V	88.6	48.9	11.3	48.0	83.7	2.3	51.1	6.7	25.0	92.1	43.5	10.3	27.3	16.0	71.0	47.6	90.3	43.5

Table G: Full results for AAD in the setting with options. We report Standard accuracy, AAD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
Standard Acc.																		
LLaVA-1.5	62.9	79.8	46.8	88.0	67.3	61.4	51.1	33.3	87.5	97.4	89.1	6.9	13.6	56.0	90.3	88.1	97.8	76.1
CogVLM	71.4	87.2	17.7	84.0	93.9	56.8	53.3	13.3	90.6	100.0	89.1	24.1	4.5	36.0	80.6	88.1	96.8	87.0
Qwen-VL-Chat	60.0	92.6	32.3	66.0	89.8	34.1	33.3	20.0	65.6	97.4	87.0	6.9	6.8	32.0	87.1	88.1	94.6	73.9
Gemini-Pro	74.3	89.4	32.3	64.0	95.9	36.4	60.0	20.0	87.5	97.4	84.8	24.1	52.3	44.0	87.1	78.6	94.6	73.9
LLaVA-NeXT-13B	68.6	84.0	50.0	86.0	87.8	61.4	60.0	53.3	93.8	100.0	89.1	27.6	22.7	60.0	93.5	88.1	97.8	87.0
LLaVA-NeXT-34B	82.9	88.3	62.9	90.0	91.8	75.0	84.4	46.7	96.9	97.4	93.5	48.3	59.1	68.0	96.8	90.5	98.9	93.5
GPT-4V	94.3	81.9	24.2	70.0	83.7	36.4	82.2	26.7	75.0	97.4	93.5	31.0	63.6	32.0	100.0	81.0	97.8	84.8
AAD Acc.																		
LLaVA-1.5	28.6	50.0	21.0	68.0	87.8	0.0	6.7	46.7	21.9	97.4	30.4	10.3	29.5	56.0	54.8	57.1	84.9	17.4
CogVLM	25.7	50.0	6.5	50.0	81.6	0.0	48.9	0.0	37.5	86.8	52.2	13.8	15.9	0.0	58.1	64.3	65.6	32.6
Qwen-VL-Chat	14.3	20.2	4.8	2.0	53.1	0.0	6.7	0.0	15.6	39.5	17.4	0.0	20.5	4.0	38.7	7.1	44.1	21.7
Gemini-Pro	34.3	71.3	32.3	52.0	87.8	13.6	33.3	6.7	31.2	78.9	32.6	31.0	2.3	20.0	64.5	50.0	77.4	34.8
LLaVA-NeXT-13B	2.9	4.3	4.8	16.0	59.2	0.0	0.0	0.0	0.0	68.4	4.3	0.0	0.0	20.0	48.4	38.1	41.9	8.7
LLaVA-NeXT-34B	17.1	35.1	9.7	22.0	75.5	4.5	6.7	13.3	6.2	78.9	8.7	6.9	4.5	16.0	64.5	31.0	69.9	10.9
GPT-4V	62.9	89.4	14.5	64.0	100.0	0.0	35.6	53.3	18.8	92.1	67.4	20.7	29.5	88.0	67.7	69.0	93.5	60.9
Dual Acc.																		
LLaVA-1.5	28.6	44.7	12.9	64.0	61.2	0.0	6.7	6.7	15.6	94.7	30.4	0.0	4.5	32.0	54.8	57.1	83.9	17.4
CogVLM	25.7	50.0	4.8	48.0	79.6	0.0	26.7	0.0	37.5	86.8	52.2	6.9	0.0	0.0	51.6	61.9	65.6	30.4
Qwen-VL-Chat	14.3	19.1	1.6	2.0	46.9	0.0	4.4	0.0	12.5	36.8	15.2	0.0	2.3	4.0	38.7	7.1	44.1	17.4
Gemini-Pro	34.3	70.2	11.3	36.0	87.8	2.3	22.2	0.0	18.8	78.9	26.1	6.9	2.3	8.0	64.5	40.5	75.3	26.1
LLaVA-NeXT-13B	2.9	4.3	4.8	16.0	59.2	0.0	0.0	0.0	0.0	68.4	4.3	0.0	0.0	8.0	48.4	38.1	41.9	8.7
LLaVA-NeXT-34B	17.1	35.1	9.7	22.0	75.5	4.5	4.4	13.3	6.2	78.9	8.7	6.9	4.5	12.0	64.5	31.0	69.9	10.9
GPT-4V	60.0	74.5	8.1	48.0	83.7	0.0	31.1	6.7	12.5	89.5	60.9	0.0	25.0	20.0	67.7	57.1	92.5	54.3

Table H: Full results for AAD in the setting with instructions. We report Standard accuracy, AAD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
Standard Acc.																		
LLaVA-1.5	62.9	81.9	45.2	90.0	63.3	59.1	48.9	26.7	81.2	94.7	89.1	6.9	13.6	44.0	87.1	88.1	96.8	69.6
CogVLM	14.3	8.5	0.0	14.0	0.0	0.0	6.7	0.0	0.0	0.0	13.0	0.0	0.0	0.0	9.7	0.0	9.7	6.5
Qwen-VL-Chat	48.6	85.1	14.5	74.0	61.2	2.3	26.7	13.3	46.9	97.4	78.3	3.4	6.8	16.0	87.1	78.6	87.1	50.0
Gemini-Pro	68.6	84.0	21.0	62.0	95.9	34.1	53.3	20.0	78.1	94.7	89.1	20.7	47.7	28.0	74.2	78.6	90.3	67.4
LLaVA-NeXT-13B	68.6	81.9	46.8	82.0	71.4	54.5	55.6	46.7	46.9	89.5	76.1	13.8	15.9	28.0	80.6	85.7	94.6	82.6
LLaVA-NeXT-34B	77.1	64.9	58.1	84.0	53.1	50.0	77.8	46.7	90.6	78.9	87.0	24.1	25.0	20.0	71.0	83.3	89.2	80.4
GPT-4V	94.3	62.8	19.4	58.0	71.4	40.9	53.3	13.3	75.0	94.7	78.3	31.0	59.1	24.0	93.5	73.8	95.7	76.1
AAD Acc.																		
LLaVA-1.5	28.6	36.2	16.1	46.0	91.8	0.0	11.1	60.0	53.1	94.7	58.7	20.7	22.7	84.0	58.1	47.6	73.1	17.4
CogVLM	91.4	93.6	75.8	94.0	100.0	93.2	95.6	40.0	93.8	100.0	93.5	93.1	90.9	96.0	96.8	100.0	92.5	95.7
Qwen-VL-Chat	17.1	41.5	16.1	36.0	71.4	15.9	6.7	26.7	53.1	71.1	47.8	41.4	38.6	64.0	51.6	28.6	57.0	37.0
Gemini-Pro	45.7	86.2	48.4	64.0	93.9	4.5	28.9	20.0	34.4	97.4	67.4	24.1	2.3	72.0	71.0	57.1	76.3	43.5
LLaVA-NeXT-13B	31.4	42.6	25.8	54.0	87.8	0.0	46.7	13.3	68.8	92.1	56.5	10.3	18.2	64.0	58.1	64.3	66.7	41.3
LLaVA-NeXT-34B	71.4	87.2	40.3	86.0	98.0	59.1	73.3	80.0	78.1	97.4	73.9	48.3	61.4	96.0	93.5	90.5	90.3	67.4
GPT-4V	80.0	96.8	74.2	90.0	100.0	36.4	88.9	73.3	40.6	97.4	93.5	62.1	43.2	100.0	80.6	97.6	95.7	87.0
Dual Acc.																		
LLaVA-1.5	28.6	33.0	9.7	46.0	59.2	0.0	11.1	13.3	43.8	89.5	58.7	3.4	2.3	36.0	54.8	47.6	72.0	17.4
CogVLM	5.7	7.4	0.0	12.0	0.0	0.0	6.7	0.0	0.0	0.0	8.7	0.0	0.0	0.0	6.5	0.0	4.3	6.5
Qwen-VL-Chat	8.6	36.2	3.2	28.0	40.8	2.3	2.2	0.0	28.1	68.4	37.0	0.0	4.5	4.0	51.6	23.8	52.7	13.0
Gemini-Pro	42.9	77.7	12.9	46.0	89.8	0.0	11.1	6.7	21.9	92.1	58.7	13.8	0.0	20.0	61.3	50.0	71.0	28.3
LLaVA-NeXT-13B	28.6	37.2	19.4	50.0	63.3	0.0	40.0	13.3	25.0	81.6	50.0	0.0	6.8	20.0	48.4	59.5	63.4	34.8
LLaVA-NeXT-34B	62.9	57.4	29.0	78.0	51.0	34.1	62.2	33.3	75.0	76.3	65.2	10.3	9.1	16.0	64.5	73.8	80.6	58.7
GPT-4V	77.1	60.6	12.9	54.0	71.4	20.5	44.4	13.3	34.4	92.1	73.9	20.7	29.5	24.0	74.2	73.8	92.5	65.2

Table I: Full results for IASD in the base setting. We report Standard accuracy, IVQD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
Standard Acc.																		
LLaVA-1.5	61.5	80.4	46.8	87.0	90.6	59.0	44.2	35.0	88.1	95.1	74.6	26.2	16.3	42.9	84.8	83.7	94.9	68.6
CogVLM	71.8	85.6	29.9	74.1	86.8	41.0	58.1	20.0	88.1	97.6	77.8	28.6	7.0	14.3	75.8	75.5	95.9	84.3
Qwen-VL-Chat	66.7	93.8	27.3	68.5	90.6	10.3	30.2	30.0	40.5	92.7	68.3	14.3	11.6	14.3	81.8	75.5	92.9	80.4
Gemini-Pro	74.4	85.6	26.0	72.2	86.8	23.1	51.2	15.0	71.4	92.7	77.8	38.1	51.2	28.6	90.9	73.5	89.8	66.7
LLaVA-NeXT-13B	69.2	83.5	50.6	83.3	88.7	59.0	55.8	50.0	59.5	100.0	65.1	21.4	16.3	25.7	87.9	77.6	95.9	84.3
LLaVA-NeXT-34B	79.5	83.5	59.7	74.1	71.7	51.3	81.4	40.0	64.3	95.1	74.6	31.0	46.5	14.3	93.9	65.3	93.9	70.6
GPT-4V	94.9	55.7	44.2	81.5	88.7	46.2	88.4	20.0	64.3	95.1	76.2	26.2	67.4	22.9	90.9	79.6	96.9	92.2
IASD Acc.																		
LLaVA-1.5	23.1	1.0	10.4	5.6	1.9	15.4	34.9	25.0	28.6	4.9	38.1	9.5	11.6	2.9	6.1	10.2	1.0	3.9
CogVLM	2.6	0.0	1.3	0.0	0.0	5.1	0.0	0.0	2.4	0.0	0.0	0.0	2.3	0.0	0.0	0.0	0.0	0.0
Qwen-VL-Chat	35.9	48.5	28.6	57.4	30.2	51.3	51.2	50.0	71.4	58.5	41.3	31.0	46.5	54.3	18.2	34.7	29.6	23.5
Gemini-Pro	38.5	72.2	40.3	42.6	30.2	53.8	37.2	25.0	57.1	61.0	50.8	33.3	39.5	31.4	54.5	46.9	41.8	37.3
LLaVA-NeXT-13B	71.8	54.6	32.5	46.3	34.0	30.8	76.7	45.0	83.3	48.8	74.6	69.0	83.7	60.0	15.2	61.2	13.3	47.1
LLaVA-NeXT-34B	79.5	89.7	70.1	87.0	73.6	74.4	100.0	80.0	100.0	92.7	82.5	76.2	86.0	85.7	54.5	87.8	59.2	98.0
GPT-4V	89.7	80.4	85.7	87.0	84.9	82.1	83.7	90.0	92.9	92.7	81.0	90.5	86.0	88.6	90.9	79.6	86.7	84.3
Dual Acc.																		
LLaVA-1.5	17.9	1.0	2.6	3.7	1.9	12.8	11.6	10.0	28.6	2.4	27.0	0.0	2.3	2.9	3.0	8.2	1.0	2.0
CogVLM	2.6	0.0	1.3	0.0	0.0	5.1	0.0	0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen-VL-Chat	28.2	47.4	11.7	40.7	28.3	2.6	11.6	15.0	26.2	53.7	25.4	7.1	7.0	2.9	12.1	30.6	27.6	17.6
Gemini-Pro	30.8	64.9	11.7	35.2	22.6	17.9	18.6	5.0	42.9	56.1	36.5	11.9	16.3	17.1	48.5	32.7	35.7	31.4
LLaVA-NeXT-13B	48.7	47.4	16.9	42.6	30.2	20.5	44.2	15.0	47.6	48.8	47.6	14.3	14.0	17.1	9.1	44.9	13.3	31.4
LLaVA-NeXT-34B	66.7	76.3	40.3	66.7	52.8	41.0	81.4	35.0	64.3	87.8	61.9	26.2	39.5	8.6	48.5	57.1	56.1	70.6
GPT-4V	84.6	46.4	39.0	74.1	75.5	35.9	76.7	20.0	57.1	87.8	61.9	23.8	62.8	17.1	84.8	59.2	84.7	80.4

Table J: Full results for IASD in the setting with options. We report Standard accuracy, IASD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
Standard Acc.																		
LLaVA-1.5	61.5	77.3	44.2	88.9	64.2	61.5	48.8	35.0	88.1	92.7	73.0	21.4	11.6	45.7	84.8	85.7	94.9	72.5
CogVLM	71.8	85.6	16.9	77.8	88.7	56.4	53.5	15.0	88.1	97.6	76.2	19.0	4.7	37.1	75.8	77.6	93.9	84.3
Qwen-VL-Chat	61.5	91.8	27.3	61.1	84.9	33.3	34.9	20.0	73.8	92.7	68.3	11.9	7.0	28.6	81.8	77.6	91.8	74.5
Gemini-Pro	76.9	88.7	27.3	64.8	90.6	35.9	62.8	20.0	83.3	97.6	74.6	40.5	53.5	37.1	84.8	71.4	92.9	72.5
LLaVA-NeXT-13B	66.7	81.4	48.1	83.3	83.0	61.5	60.5	50.0	92.9	97.6	74.6	31.0	23.3	48.6	90.9	81.6	95.9	86.3
LLaVA-NeXT-34B	79.5	84.5	57.1	87.0	86.8	74.4	86.0	45.0	95.2	97.6	82.5	45.2	58.1	54.3	93.9	83.7	96.9	94.1
GPT-4V	89.7	80.4	27.3	70.4	81.1	33.3	83.7	25.0	73.8	95.1	81.0	28.6	65.1	25.7	97.0	71.4	95.9	84.3
IASD Acc.																		
LLaVA-1.5	66.7	74.2	63.6	72.2	58.5	66.7	86.0	65.0	73.8	78.0	65.1	71.4	62.8	42.9	48.5	73.5	60.2	66.7
CogVLM	17.9	23.7	28.6	25.9	20.8	5.1	25.6	25.0	52.4	22.0	38.1	19.0	0.0	31.4	24.2	34.7	19.4	27.5
Qwen-VL-Chat	46.2	48.5	42.9	50.0	30.2	35.9	53.5	55.0	59.5	41.5	66.7	26.2	60.5	14.3	39.4	57.1	38.8	41.2
Gemini-Pro	59.0	73.2	59.7	70.4	60.4	59.0	74.4	55.0	71.4	68.3	74.6	59.5	37.2	51.4	81.8	73.5	72.4	86.3
LLaVA-NeXT-13B	35.9	43.3	31.2	24.1	35.8	15.4	48.8	45.0	59.5	43.9	34.9	42.9	30.2	31.4	33.3	59.2	38.8	56.9
LLaVA-NeXT-34B	25.6	20.6	18.2	29.6	35.8	12.8	46.5	35.0	7.1	14.6	44.4	47.6	18.6	22.9	36.4	16.3	39.8	33.3
GPT-4V	94.9	92.8	93.5	90.7	96.2	97.4	90.7	100.0	95.2	92.7	90.5	95.2	95.3	91.4	97.0	91.8	95.9	96.1
Dual Acc.																		
LLaVA-1.5	43.6	58.8	32.5	64.8	35.8	41.0	39.5	20.0	61.9	78.0	49.2	19.0	4.7	25.7	42.4	59.2	57.1	52.9
CogVLM	17.9	22.7	5.2	18.5	20.8	2.6	16.3	5.0	47.6	22.0	25.4	2.4	0.0	20.0	21.2	28.6	18.4	25.5
Qwen-VL-Chat	35.9	45.4	7.8	33.3	26.4	17.9	16.3	5.0	42.9	36.6	46.0	2.4	4.7	5.7	30.3	44.9	36.7	29.4
Gemini-Pro	53.8	64.9	14.3	51.9	54.7	25.6	48.8	10.0	57.1	65.9	52.4	26.2	20.9	28.6	66.7	51.0	69.4	64.7
LLaVA-NeXT-13B	33.3	34.0	19.5	20.4	28.3	12.8	27.9	20.0	57.1	43.9	25.4	14.3	4.7	17.1	33.3	46.9	37.8	45.1
LLaVA-NeXT-34B	20.5	16.5	13.0	24.1	32.1	10.3	37.2	15.0	7.1	12.2	34.9	21.4	9.3	14.3	36.4	12.2	38.8	33.3
GPT-4V	84.6	74.2	26.0	64.8	79.2	30.8	76.7	25.0	69.0	87.8	73.0	26.2	65.1	20.0	93.9	63.3	92.9	80.4

Table K: Full results for IASD in the setting with instructions. We report Standard accuracy, IASD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
Standard Acc.																		
LLaVA-1.5	61.5	79.4	41.6	88.9	58.5	61.5	46.5	25.0	83.3	90.2	73.0	21.4	11.6	37.1	81.8	85.7	93.9	70.6
CogVLM	12.8	7.2	0.0	13.0	0.0	0.0	7.0	0.0	0.0	0.0	9.5	0.0	0.0	0.0	6.1	0.0	12.2	5.9
Qwen-VL-Chat	51.3	81.4	15.6	66.7	58.5	2.6	23.3	15.0	47.6	90.2	60.3	2.4	7.0	14.3	81.8	67.3	82.7	54.9
Gemini-Pro	66.7	83.5	19.5	63.0	88.7	33.3	55.8	25.0	73.8	92.7	76.2	35.7	51.2	22.9	72.7	71.4	88.8	62.7
LLaVA-NeXT-13B	66.7	79.4	44.2	77.8	67.9	59.0	58.1	40.0	47.6	87.8	66.7	19.0	23.3	22.9	75.8	77.6	92.9	82.4
LLaVA-NeXT-34B	74.4	61.9	54.5	81.5	50.9	46.2	76.7	45.0	90.5	78.0	73.0	19.0	23.3	17.1	69.7	75.5	87.8	80.4
GPT-4V	89.7	61.9	20.8	59.3	67.9	41.0	55.8	15.0	73.8	92.7	65.1	33.3	60.5	20.0	90.9	65.3	93.9	74.5
IASD Acc.																		
LLaVA-1.5	82.1	72.2	87.0	74.1	73.6	74.4	95.3	90.0	95.2	85.4	84.1	81.0	86.0	77.1	75.8	79.6	74.5	76.5
CogVLM	97.4	97.9	89.6	92.6	92.5	92.3	95.3	80.0	97.6	97.6	96.8	71.4	62.8	85.7	100.0	98.0	93.9	100.0
Qwen-VL-Chat	79.5	58.8	48.1	51.9	54.7	69.2	60.5	65.0	85.7	51.2	60.3	69.0	58.1	68.6	45.5	44.9	43.9	47.1
Gemini-Pro	71.8	88.7	75.3	75.9	73.6	84.6	88.4	75.0	90.5	87.8	82.5	69.0	79.1	68.6	90.9	79.6	75.5	92.2
LLaVA-NeXT-13B	74.4	85.6	92.2	75.9	90.6	84.6	97.7	95.0	97.6	87.8	93.7	83.3	86.0	94.3	87.9	87.8	86.7	98.0
LLaVA-NeXT-34B	94.9	97.9	94.8	100.0	94.3	100.0	97.7	100.0	97.6	97.6	98.4	97.6	97.7	97.1	97.0	98.0	94.9	100.0
GPT-4V	100.0	97.9	96.1	96.3	100.0	100.0	97.7	100.0	95.2	97.6	100.0	97.6	97.7	100.0	100.0	95.9	98.0	98.0
Dual Acc.																		
LLaVA-1.5	51.3	57.7	33.8	68.5	39.6	51.3	41.9	20.0	78.6	80.5	63.5	19.0	7.0	31.4	60.6	65.3	69.4	56.9
CogVLM	12.8	7.2	0.0	11.1	0.0	0.0	7.0	0.0	0.0	0.0	7.9	0.0	0.0	0.0	6.1	0.0	9.2	5.9
Qwen-VL-Chat	43.6	47.4	9.1	35.2	30.2	2.6	16.3	5.0	40.5	43.9	38.1	2.4	4.7	8.6	39.4	26.5	33.7	19.6
Gemini-Pro	53.8	76.3	14.3	48.1	62.3	30.8	53.5	20.0	66.7	82.9	65.1	23.8	46.5	20.0	66.7	61.2	68.4	58.8
LLaVA-NeXT-13B	61.5	66.0	40.3	64.8	62.3	56.4	55.8	35.0	47.6	78.0	63.5	16.7	20.9	22.9	69.7	65.3	80.6	80.4
LLaVA-NeXT-34B	71.8	59.8	51.9	81.5	45.3	46.2	74.4	45.0	88.1	75.6	71.4	16.7	23.3	17.1	66.7	73.5	82.7	80.4
GPT-4V	89.7	59.8	19.5	55.6	67.9	41.0	55.8	15.0	69.0	90.2	65.1	31.0	60.5	20.0	90.9	63.3	91.8	72.5

Table L: Full results for IVQD in the base setting. We report Standard accuracy, IASD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#11	#12	#17
Standard Acc.												
LLaVA-1.5	54.8	88.2	58.3	72.2	92.9	82.6	46.7	26.7	88.4	68.8	13.0	87.5
CogVLM	67.7	89.7	36.1	55.6	85.7	43.5	57.8	13.3	88.4	93.8	8.7	91.7
Qwen-VL-Chat	61.3	97.1	22.2	38.9	92.9	60.9	40.0	20.0	46.5	68.8	17.4	83.3
Gemini-Pro	71.0	94.1	11.1	44.4	92.9	65.2	60.0	13.3	72.1	93.8	34.8	87.5
LLaVA-NeXT-13B	64.5	86.8	55.6	72.2	92.9	82.6	60.0	46.7	65.1	75.0	17.4	87.5
LLaVA-NeXT-34B	74.2	92.6	58.3	72.2	71.4	78.3	82.2	33.3	69.8	93.8	39.1	83.3
GPT-4V	93.5	45.6	50.0	55.6	78.6	78.3	93.3	13.3	65.1	75.0	26.1	91.7
IVQD Acc.												
LLaVA-1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CogVLM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen-VL-Chat	3.2	41.2	77.8	38.9	14.3	43.5	48.9	40.0	76.7	6.2	26.1	41.7
Gemini-Pro	9.7	19.1	66.7	33.3	28.6	21.7	33.3	26.7	55.8	18.8	0.0	25.0
LLaVA-NeXT-13B	22.6	75.0	25.0	11.1	42.9	0.0	48.9	20.0	79.1	0.0	21.7	25.0
LLaVA-NeXT-34B	45.2	89.7	69.4	38.9	71.4	21.7	91.1	53.3	95.3	37.5	43.5	62.5
GPT-4V	96.8	98.5	100.0	88.9	100.0	91.3	97.8	100.0	100.0	93.8	91.3	95.8
Dual Acc.												
LLaVA-1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CogVLM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen-VL-Chat	3.2	41.2	11.1	11.1	7.1	17.4	17.8	13.3	34.9	6.2	4.3	41.7
Gemini-Pro	9.7	19.1	2.8	22.2	28.6	17.4	20.0	0.0	48.8	18.8	0.0	20.8
LLaVA-NeXT-13B	19.4	64.7	19.4	11.1	35.7	0.0	28.9	20.0	46.5	0.0	0.0	25.0
LLaVA-NeXT-34B	45.2	85.3	30.6	22.2	57.1	21.7	77.8	26.7	65.1	37.5	17.4	54.2
GPT-4V	90.3	45.6	50.0	50.0	78.6	73.9	91.1	13.3	65.1	75.0	17.4	87.5

Table M: Full results for IVQD in the setting with options. We report Standard accuracy, IVQD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#11	#12	#17
Standard Acc.												
LLaVA-1.5	51.6	88.2	61.1	83.3	85.7	82.6	51.1	26.7	90.7	68.8	8.7	91.7
CogVLM	67.7	89.7	16.7	50.0	85.7	87.0	60.0	6.7	90.7	81.2	8.7	91.7
Qwen-VL-Chat	54.8	92.6	22.2	44.4	92.9	73.9	28.9	6.7	74.4	56.2	8.7	83.3
Gemini-Pro	80.6	94.1	11.1	50.0	92.9	78.3	60.0	6.7	86.0	87.5	34.8	83.3
LLaVA-NeXT-13B	61.3	85.3	58.3	72.2	85.7	87.0	57.8	40.0	93.0	75.0	17.4	87.5
LLaVA-NeXT-34B	77.4	94.1	58.3	77.8	78.6	87.0	84.4	33.3	95.3	100.0	52.2	91.7
GPT-4V	93.5	69.1	19.4	50.0	85.7	73.9	57.8	13.3	69.8	87.5	30.4	91.7
IVQD Acc.												
LLaVA-1.5	29.0	95.6	47.2	11.1	71.4	0.0	15.6	26.7	76.7	6.2	4.3	66.7
CogVLM	0.0	41.2	0.0	0.0	0.0	0.0	42.2	0.0	60.5	6.2	0.0	37.5
Qwen-VL-Chat	16.1	55.9	80.6	38.9	57.1	17.4	13.3	13.3	74.4	31.2	0.0	50.0
Gemini-Pro	48.4	91.2	91.7	83.3	100.0	60.9	93.3	60.0	97.7	68.8	47.8	87.5
LLaVA-NeXT-13B	29.0	92.6	83.3	16.7	85.7	0.0	13.3	26.7	48.8	6.2	0.0	62.5
LLaVA-NeXT-34B	35.5	97.1	75.0	22.2	64.3	21.7	55.6	40.0	79.1	12.5	4.3	62.5
GPT-4V	90.3	100.0	100.0	94.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Dual Acc.												
LLaVA-1.5	22.6	85.3	30.6	11.1	71.4	0.0	13.3	6.7	67.4	0.0	0.0	66.7
CogVLM	0.0	36.8	0.0	0.0	0.0	0.0	20.0	0.0	58.1	6.2	0.0	37.5
Qwen-VL-Chat	9.7	55.9	19.4	16.7	50.0	17.4	6.7	0.0	58.1	25.0	0.0	45.8
Gemini-Pro	45.2	85.3	11.1	50.0	92.9	52.2	55.6	6.7	86.0	62.5	8.7	79.2
LLaVA-NeXT-13B	25.8	79.4	50.0	16.7	71.4	0.0	8.9	20.0	46.5	6.2	0.0	58.3
LLaVA-NeXT-34B	35.5	92.6	41.7	16.7	57.1	21.7	46.7	26.7	74.4	12.5	4.3	62.5
GPT-4V	83.9	69.1	19.4	50.0	85.7	73.9	57.8	13.3	69.8	87.5	30.4	91.7

Table N: Full results for IVQD in the setting with instructions. We report Standard accuracy, IVQD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#11	#12	#17
Standard Acc.												
LLaVA-1.5	51.6	88.2	58.3	77.8	92.9	82.6	48.9	20.0	90.7	62.5	8.7	91.7
CogVLM	71.0	88.2	30.6	44.4	85.7	56.5	57.8	13.3	93.0	81.2	8.7	91.7
Qwen-VL-Chat	45.2	80.9	13.9	33.3	35.7	30.4	24.4	13.3	41.9	56.2	8.7	66.7
Gemini-Pro	67.7	92.6	11.1	38.9	92.9	56.5	44.4	13.3	79.1	87.5	26.1	83.3
LLaVA-NeXT-13B	58.1	86.8	58.3	66.7	78.6	87.0	55.6	33.3	83.7	75.0	17.4	87.5
LLaVA-NeXT-34B	71.0	94.1	58.3	77.8	85.7	82.6	84.4	33.3	93.0	93.8	39.1	91.7
GPT-4V	90.3	79.4	19.4	33.3	71.4	60.9	44.4	13.3	67.4	75.0	17.4	83.3
IVQD Acc.												
LLaVA-1.5	22.6	80.9	19.4	16.7	85.7	0.0	8.9	33.3	53.5	18.8	4.3	54.2
CogVLM	0.0	35.3	0.0	0.0	0.0	0.0	2.2	0.0	16.3	6.2	0.0	4.2
Qwen-VL-Chat	51.6	75.0	72.2	55.6	57.1	52.2	80.0	80.0	95.3	37.5	82.6	70.8
Gemini-Pro	96.8	100.0	100.0	100.0	100.0	95.7	100.0	86.7	100.0	93.8	82.6	100.0
LLaVA-NeXT-13B	58.1	97.1	88.9	44.4	92.9	0.0	91.1	46.7	95.3	50.0	26.1	66.7
LLaVA-NeXT-34B	90.3	100.0	100.0	61.1	100.0	52.2	100.0	100.0	100.0	93.8	69.6	95.8
GPT-4V	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Dual Acc.												
LLaVA-1.5	16.1	76.5	11.1	16.7	78.6	0.0	6.7	6.7	44.2	12.5	0.0	54.2
CogVLM	0.0	33.8	0.0	0.0	0.0	0.0	0.0	0.0	16.3	6.2	0.0	4.2
Qwen-VL-Chat	22.6	63.2	8.3	11.1	14.3	8.7	20.0	13.3	39.5	18.8	4.3	50.0
Gemini-Pro	67.7	92.6	11.1	38.9	92.9	56.5	44.4	13.3	79.1	87.5	13.0	83.3
LLaVA-NeXT-13B	51.6	85.3	52.8	38.9	71.4	0.0	53.3	20.0	79.1	43.8	4.3	58.3
LLaVA-NeXT-34B	67.7	94.1	58.3	44.4	85.7	43.5	84.4	33.3	93.0	87.5	17.4	87.5
GPT-4V	90.3	79.4	19.4	33.3	71.4	60.9	44.4	13.3	67.4	75.0	17.4	83.3

Table O: Full results for AAD with instruction tuning. We report Standard accuracy, AAD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
Standard Acc.																		
LLaVA-NeXT-13B	68.6	76.6	50.0	88.0	73.5	56.8	60.0	60.0	81.2	86.8	76.1	13.8	17.8	52.0	80.6	85.7	92.5	73.9
LLaVA-NeXT-34B	80.0	78.7	61.3	88.0	83.7	70.5	82.2	40.0	93.8	92.1	80.4	24.1	57.8	48.0	93.5	90.5	96.8	89.1
AAD Acc.																		
LLaVA-NeXT-13B	45.7	68.1	21.0	54.0	89.8	0.0	31.1	6.7	53.1	97.4	84.8	34.5	22.2	88.0	74.2	61.9	93.5	65.2
LLaVA-NeXT-34B	65.7	84.0	51.6	80.0	100.0	25.0	55.6	33.3	96.9	100.0	93.5	31.0	24.4	84.0	96.8	83.3	97.8	58.7
Dual Acc.																		
LLaVA-NeXT-13B	42.9	57.4	12.9	52.0	67.3	0.0	17.8	6.7	46.9	84.2	69.6	3.4	13.3	40.0	61.3	59.5	88.2	52.2
LLaVA-NeXT-34B	62.9	69.1	43.5	76.0	83.7	18.2	53.3	20.0	90.6	92.1	78.3	6.9	20.0	40.0	93.5	76.2	95.7	54.3

Table P: Full results for IASD with instruction tuning. We report Standard accuracy, IASD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18
Standard Acc.																		
LLaVA-NeXT-13B	66.7	74.2	46.8	81.5	67.9	56.4	58.1	55.0	81.0	85.4	63.5	23.8	18.2	42.9	75.8	75.5	90.8	74.5
LLaVA-NeXT-34B	76.9	76.3	57.1	85.2	79.2	69.2	81.4	40.0	92.9	92.7	71.4	28.6	56.8	37.1	87.9	83.7	95.9	88.2
IASD Acc.																		
LLaVA-NeXT-13B	97.4	91.8	90.9	85.2	92.5	61.5	97.7	95.0	90.5	92.7	95.2	90.5	95.5	100.0	90.9	89.8	89.8	98.0
LLaVA-NeXT-34B	100.0	96.9	96.1	96.3	98.1	100.0	97.7	100.0	97.6	100.0	96.8	100.0	97.7	97.1	100.0	95.9	98.0	100.0
Dual Acc.																		
LLaVA-NeXT-13B	66.7	66.0	42.9	74.1	64.2	30.8	58.1	50.0	73.8	80.5	63.5	21.4	15.9	42.9	72.7	65.3	81.6	72.5
LLaVA-NeXT-34B	76.9	74.2	55.8	83.3	77.4	69.2	79.1	40.0	90.5	92.7	68.3	28.6	56.8	37.1	87.9	79.6	93.9	88.2

Table Q: Full results for IVQD with instruction tuning. We report Standard accuracy, IVQD accuracy, and Dual accuracy.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#11	#12	#17
Standard Acc.												
LLaVA-NeXT-13B	61.3	82.4	52.8	72.2	64.3	82.6	60.0	46.7	81.4	75.0	13.0	87.5
LLaVA-NeXT-34B	74.2	82.4	52.8	72.2	78.6	69.6	82.2	33.3	93.0	87.5	34.8	95.8
IVQD Acc.												
LLaVA-NeXT-13B	87.1	100.0	100.0	83.3	100.0	4.3	100.0	73.3	100.0	81.2	69.6	75.0
LLaVA-NeXT-34B	90.3	100.0	100.0	100.0	100.0	78.3	100.0	93.3	100.0	93.8	87.0	83.3
Dual Acc.												
LLaVA-NeXT-13B	61.3	82.4	52.8	66.7	64.3	0.0	60.0	33.3	81.4	68.8	8.7	70.8
LLaVA-NeXT-34B	64.5	82.4	52.8	72.2	78.6	47.8	82.2	33.3	93.0	81.2	26.1	79.2