Aligning "Hallucinations": Benchmarking LLMs and VLMs with Humans on Blind Visual Question Answering

Anonymous ACL submission

Abstract

Pretrained linguistic knowledge provides an essential semantic foundation for modern Vision-Language Models (VLMs), but its impact on genuine visual grounding remains unclear. We introduce "hallucination alignment", a framework that for the first time systematically compares models and human responses on visual question answering (VQA) *without visual input*. To this end, we gather the first large human dataset on blinded VQA-derived questions, evaluating Large Language Models (LLMs) and VLMs against human performance and answer patterns. We find that linguistic priors in VLMs enable blind performance exceeding both LLMs and humans. Likewise, answer patterns for both LLMs and VLMs differ significantly from human answers. We show that it is possible to align VLMs to human blinded answers at no cost to visually-grounded inference, creating better aligned multimodal models.

1 Introduction and Related Work

Vision–Language Models (VLMs) (Radford et al., 2021; Huang et al., 2023a,b; Liu et al., 2023a; Zhu et al., 2025; Hurst et al., 2024; Bai et al., 2025) inherit strong linguistic priors from their foundational Large Language Model (LLM) components (Touvron et al., 2023). This enables impressive fluency in answering visually grounded questions (Antol et al., 2015; Goyal et al., 2016; Wu et al., 2023; Liu et al., 2023b; Li et al., 2023; Fu et al., 2023; Chen et al., 2024). However, aligning these linguistic expectations with visual evidence remains a significant challenge. Their mismatch often leads to the phenomenon of *hallucination*, where models confidently produce plausible responses as they heavily rely on spurious biases or linguistic cues over visual information (Agrawal et al., 2017; Xiao and Wang, 2021a; Ji et al., 2023; Xiao and Wang, 2021b; Huang et al., 2024; Han et al., 2024; Agrawal et al., 2017; Rohrbach et al., 2018)—likely since the training objectives incentivize probable linguistic responses over explicit visual grounding (Kalai et al., 2025; Ye et al., 2024; Lee et al., 2024).

Previous work Chen et al. (2024) has pointed out that models lack proper visual grounding, demonstrating how VLMs often succeed in VQA tasks without visual input. They designed a benchmark, MMStar, that aims to measure visual grounding and logical reasoning, via categories such as coarse and fine-grained perception and instance reasoning. Likewise, human cognition also relies deeply on linguistic and conceptual priors when sensory information is absent or ambiguous (Tversky and Kahneman, 1974). Importantly, however, while models are tested broadly across benchmarks, direct empirical comparisons between **human and VLM behavior** on such "hallucinations" are scarce.

To bridge this gap, we collect human data on this behavior in 'blind' scenarios providing only the probing questions from VQA benchmarks and evaluate VLMs and LLMs against human performance. We collect the first human-annotated blind-inference dataset from open-ended and multiple-choice VQA benchmarks including VQA 2.0 (Goyal et al., 2016) and MMStar (Chen et al., 2024), spanning different question categories. We measure the degree of hallucination alignment both in terms of accuracy but also in terms of answer patterns to provide a deeper understanding of their language priors. Finally, we investigate how these language priors interact with visual grounding abilities on the state-of-the-art VLMs by steering their bias towards human prior via fine-tuning and see how or whether it reshapes their hallucination behavior and visual grounding.

2 Experimental Setup

2.1 Blind Inference Protocol

Datasets and Models. Our hallucination evaluation set consists of 374 questions sampled across

different question types from the VQA 2.0 validation split (Goyal et al., 2016), and 267 questions from MMStar (Chen et al., 2024), drawn from four categories and excluding the Science & Technology category, as our analysis focuses on visual reasoning. We use the same test set for humans for the quadrant analysis in Section 2.1, and use another subset of 5,000 images of the VQA dataset in Section 3.3. We benchmark a set of open-source state-of-the-art LLMs and VLMs, including Qwen3 series (Bai et al., 2025; Yang et al., 2025), InternVL3.5 (Zhu et al., 2025) and LLaVA models (Liu et al., 2023a, 2024) with two language backbone variants, including Mistral (Jiang et al., 2023) and Vicuna (Zheng et al., 2023).

Human Data Collection. Using an IRB-approved protocol with standard compensation, we recruited N=20 human participants, who were instructed to answer each of the 641 VQA-like questions based solely on the textual prompt, ensuring consistency across responses. Participants were instructed to rely only on general world knowledge and linguistic cues. After answering, participants reported their confidence on a 5-point Likert scale (1–5), with higher values indicating greater confidence.

Accuracy Metrics. We report three metrics: (1) accuracy following the standard evaluation protocol in Antol et al., 2015 for VQA or percentage correct for the 4-alternative forced-choice MMStar, (2) semantic similarity for non-dichotomous or non-numeric VQA questions between predicted and ground-truth answers using Sentence-BERT embeddings (MiniLM-L6-v2; Wang et al., 2020) via cosine similarity, and for VLM models also (3) *Multi-modal Gain* (MG), which quantifies the performance drop when visual input is removed.

We evaluate the VLM models in two inference settings, 1) given blank black images with additional instruction to guess based on common world knowledge (refer to Appendix E.1 for details on instruction) and 2) original image with questions to compute multimodal gain (MG). Following Chen et al. (2024), MG is defined as $\mathrm{MG} = S_{\text{vision+text}} - S_{\text{text-only}}$, where $S_{\text{vision+text}}$ and $S_{\text{text-only}}$ denote model performance with full visual input and with a blank image, respectively.

Answer pattern alignment. We estimate answer agreement of models with humans or humans with humans via the pairwise Spearman correlation (ρ) on the list of scores of questions of each subject. Human–human agreement is computed via a leave-one-out protocol, correlating each annotator's correctness with the mean correctness of all other annotators per question, while model–human agreement correlates model correctness with aggregated human consensus; Spearman correlations are averaged directly. Confidence intervals for human and LLM baselines are estimated via leave-one-rater-out jackknife, treating raters (humans or models) as the unit of independence, where removing a rater removes all correlations involving that rater.

2.2 Fine-tuning to Human Data

To isolate the effect of alignment to human language priors, we fine-tune VLMs of 4B and 7-8B on blind human annotations (N=10 and 15) using LoRA adapters, while freezing the vision encoder and alignment modules. We leverage aggregated human confidence scores under two training objectives: (1) standard supervised fine-tuning using a single representative answer selected based on highest confidence and frequency (SFT), and (2) distributional alignment using a Jensen–Shannon divergence loss that incorporates token-level supervision derived from human confidence scores (JS). Full dataset curation and training implementation details are provided in Appendix E.

3 Experimental results

3.1 Language priors

The blind performance of humans and pretrained models, with Qwen 3 LLM as our baseline models is reported in Table 1.

Humans vs LLMs vs VLMs. Human VQA performance on dichotomous yes/no questions is better than chance, whereas guessing the exact number of items blindly proves harder. Performance on "other" question types in terms of exact match is at low, yet non-zero levels, but as can be seen, humans are able to provide similar answers to the ground truth in many cases. Performance on the more open-ended MMStar questions is similar across categories and at the reported random choice level of $\approx 25\%$(Chen et al., 2024) testifying to the dataset's design choice of testing pure visual grounding.

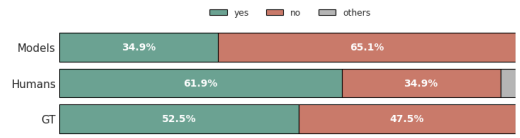
In terms of VQA performance, interestingly, VLMs show higher average accuracy in the blind condition than humans, whereas the LLM baseline

Model	LLM	Size	VQA				MMStar				
			Y/N	Num	Other	Avg	CP	FP	IR	LR	Avg
Humans (N=20)	–	–	66.6	13.8	13.0 / 45.5	39.0	28.2	26.3	28.4	27.7	27.4
LLM Baseline											
Qwen3	Qwen3	4B	54.9	19.0	24.7 / 55.2	38.8	25.0	33.7	23.6	40.9	29.7
		8B	50.3	17.1	22.1 / 53.2	35.3	17.6	29.7	20.0	59.1	26.8
Multi-Modal models											
Qwen3-VL	Qwen3	2B	71.8	17.1	23.8 / 55.5	46.4	30.9	28.7	21.8	22.7	27.2
		4B	70.3	19.0	**28.9 / 56.4**	48.0	17.6	**34.7**	25.5	40.9	28.5
		8B	71.5	21.9	27.0 / 54.9	48.0	30.9	30.7	**30.9**	50.0	32.5
InternVL 3.5	Qwen3-0.6B	1B	68.3	**38.1**	20.0 / 52.8	45.1	25.0	29.7	29.1	27.3	28.0
	Qwen3-1.7B	2B	**74.6**	17.1	19.2 / 53.0	45.8	20.6	25.7	18.2	36.4	23.6
	Qwen3-4B	4B	74.4	17.1	24.0 / 55.7	47.8	20.6	25.7	25.5	31.8	24.8
	Qwen3-8B	8B	70.3	22.9	22.8 / 54.3	45.8	27.9	32.7	27.3	**63.6**	**32.9**
LLaVA v1.5	Vicuna 1.5	7B	70.3	17.1	19.4 / 48.4	43.8	**36.8**	27.7	29.1	27.3	30.5
		7B	74.4	20.0	27.2 / 56.2	**49.4**	32.4	20.8	**30.9**	18.2	26.0
LLaVA v1.6	Mistral	7B	69.6	20.0	27.0 / 55.0	47.0	25.0	24.8	21.8	31.8	24.8

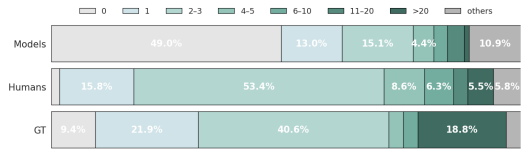
Table 1: Zero-shot blind performance of humans and models across dataset categories. Metrics for Other answer types in VQA are Accuracy/Similarity pairs; others are accuracy only. MMStar accesses include CP (coarse perception), FP (fine-grained perception), IR(instance reasoning), LR (logical reasoning) core capabilities. Bold denotes best; underline denotes second best.

performs on average similarly. As the table entries demonstrate, however, both LLMs and VLMs show different performance patterns according to question types against the human benchmark. On MMStar, several VLMs in addition show high blind performance on logical reasoning tasks, indicating successful, yet non-human-like hallucination.

Effects of Model size. Increasing the model size for LLM models shows slight decreases in performance on the blind set. Conversely, both QwenVL and InternVL have clear gains for the MMStar benchmarks with increasing model size, especially in the logical reasoning tasks. Performance on VQA shows less variability with model size. This pattern suggests that larger models encode stronger and more structured language priors.



(a) Yes or No



(b) Number

Figure 1: Responses for dichotomous/numerical questions.

3.2 Human–Model Agreement

Answer Biases. Figure 1a records answer distributions for yes/no and numerical questions in the VQA task. Here, models overall produce different answer patterns compared to humans: fewer positive answers for categorical questions, and a large bias towards answering "0" while at the same time producing significantly fewer answers in the small number regime.

Correct/incorrect matrix. To further characterize model behavior, we stratify questions based on blind correctness for both humans and models, then analyze how models behave when visual information becomes available via accuracy-based Multimodal Gain (MG). This is done for four correctness regimes (Shared Wrong, Human-Only, Shared Correct, Model-Only) using a 50% average accuracy threshold for humans and models.

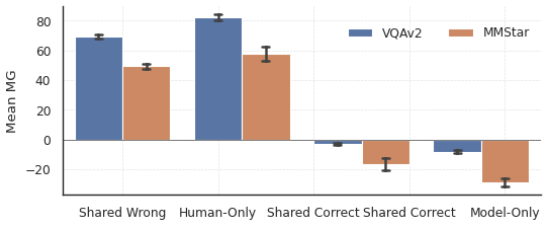


Figure 2: Multimodal gain (MG) for accuracy matrix.

When blind linguistic inference fails, adding visual input yields consistently large positive gains, as MG is largest in the *Shared Wrong* and *Human-Only* regimes as shown in Figure 2. A representative example from VQA questions yielding largest MG and high blind performance in humans, asks for the number of clocks on a tower. Humans correctly answer "1" by leveraging bias anchored by the question, whereas models default to null or generic answers as reflected in Figure 1.

In contrast, when blind correctness is already achieved, visual input contributes little or may even degrade performance, as MG collapses in

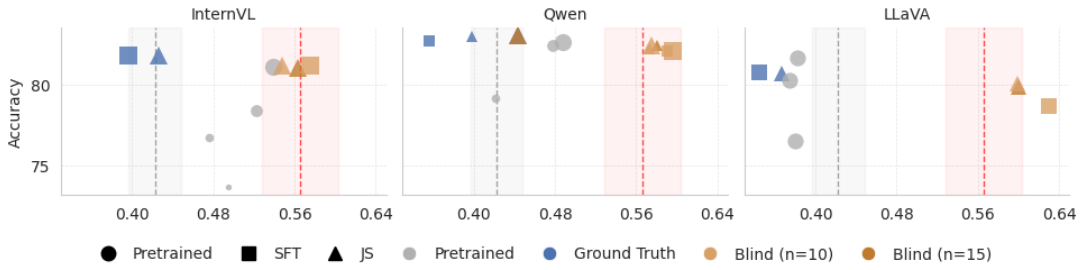


Figure 3: Human–model alignment versus test accuracy on VQA dataset. The x-axis shows Spearman correlation (ρ) between model and human correctness. Marker size=size of the models. Blue markers indicate ground truth training, orange markers are prior aligned models in different training strategies. Human baseline is highlighted in red and Qwen LLM baseline is in gray.

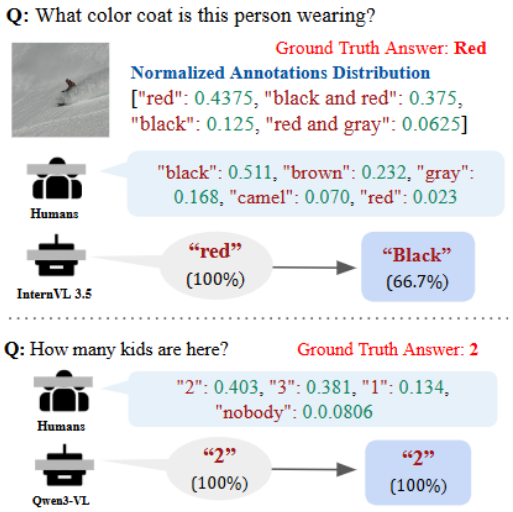


Figure 4: Answer change of InternVL 3.5 and Qwen3-VL after human responses alignment when pretrained models are correct in the blind setting.

the *Shared Correct* and *Model-Only* regime, with more pronounced degradation when humans are not correct — suggesting the blind model correctness is often driven by idiosyncratic or spurious linguistic correlations that conflict with visual evidence. Cases where visually grounded predictions diverge from heterogeneous or misleading human priors include in the question “What type of establishment are these people likely seated?”, all VLMs correctly answer *restaurant* based on probable imagined cues. In contrast, human responses include variable answers (such as playground, nursing home, park bench, school, or waiting room), reflecting reliance on ambiguous or incorrect priors in the absence of visual input.

3.3 Human–Model Prior Alignment

Models initially exhibit low alignment with human responses ($\rho = 0.455$), lower than the human baseline ($\rho = 0.566$) and closer to their LLM baseline ($\rho = 0.423$), as shown in Figure 3. However, after fine-tuning with human prior biases, alignment improves substantially ($\rho = 0.598$ for SFT and $\rho = 0.547$ for JS), while preserving visually grounded performance, incurring little to no loss in original VQA accuracy. After alignment finetuning, the models produce more generic responses similar to humans, as illustrated by the VQA examples in Figure 4 that capture the human blind-answer distribution correctly. We also note that fine-tuning on the original VQA responses does not produce better human blind-response alignment, as shown by the blue markers in Figure 3.

Qualitative examples from MMStar further highlight how the alignment could suppress the outstanding blind performance observed in 8B VLMs. In one example from the logical reasoning category estimating the proportion of an image occupied by a bus, about half of humans converge on the ground truth option (0.6). However the Qwen3-VL model regressed to other choices after the alignment, despite having the answer correct initially. Conversely, the opposite effect was observed in another example from MMStar, where the question asks "Which mood does this image convey?". More than half of the humans preferred to answer "cozy" (the ground truth answer), while the models anchored to the option "sad" before alignment. Our gathered human responses helped to compensate the degraded performance of models from incorrect but generic human priors.

4 Conclusion

In this work, we analyzed hallucination in vision–language models by comparing their language-only inference with human responses under missing visual input. We showed that VLMs outperform both humans and LLMs, with both LLMs and VLMs exhibiting significantly different answer patterns as well—a large misalignment. Using inexpensive fine-tuning, we managed to increase alignment while keeping the multimodal, visually-grounded performance intact.

Limitations and Future Work. Our models and alignment methods can further reduce leakage to existing benchmarks by better leveraging on human priors and improving generalizability, without harming the accuracy. However, our analyses rely on a limited set of model families and human annotations evaluated question subsets. Consequently, the reported correlations should be interpreted directionally rather than as precise effect sizes or causal estimates. In addition, model inference is deterministic in our evaluation, with zero-shot models evaluated using a single inference pass without task-specific optimization; as a result, individual measurements may not fully reflect variability.

Future work would benefit from scaling human annotation efforts on such human priors, with expanding the range of pretrained and fine-tuned models evaluated, and exploring explicit fine-tuning strategies that better balance the trade-off between leveraging linguistic priors and ensuring robust visual grounding. Finally, extending evaluation to larger datasets and benchmarks that explicitly probe spurious correlations or counterfactual reasoning would strengthen causal claims about the role of language priors in vision–language model behavior.

Acknowledgments

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *arXiv.org*.

Yash Goyal, Tejas Khot, D. Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *International Journal of Computer Vision*.

Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. 2024. The instinctive bias: Spurious images lead to illusion in mllms. *Preprint*, arXiv:2402.03757.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *Preprint*, arXiv:2311.17911.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023a. Language is not all you need: Aligning perception with language models. *Preprint*, arXiv:2302.14045.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, and 1 others. 2023b. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. Gpt-4o system card. In *arXiv.org*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

A. Kalai, Ofir Nachum, S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate.

Kangil Lee, Minbeom Kim, Seunghyun Yoon, Minsu Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2024. Vlind-bench: Measuring language priors in large vision-language models. In *arXiv.org*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *arXiv.org*.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*.

OpenAI. 2025. GPT-4o: Introducing GPT-4o, our new flagship model. `https://openai.com/index/introducing-4o-image-generation`. Accessed: [Current Date, e.g., 2024-05-15].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations*.

Yijun Xiao and W. Wang. 2021a. On hallucination and predictive uncertainty in conditional language generation. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Yijun Xiao and William Yang Wang. 2021b. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. 2024. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix

A Human Study Protocol and Participant Details

A total of 24 participants (9 men and 15 women; age range: 20–43 years) were included in the study. Participants were recruited via online student community forums at universities. All procedures were conducted under an IRB-approved protocol and adhered to the principles of the Declaration of Helsinki. Informed consent was obtained from all participants prior to participation.

Each participant completed 641 VQA-style questions in a single session, which took approximately one hour on average. Participants were compensated at a rate of 10,000 KRW per hour, consistent with standard compensation for similar annotation tasks.

For VQA, participants were instructed to provide free-form answers that were as concise as possible. Responses could be given in either Korean or English and were translated and processed using the same answer normalization and aggregation strategy described in Appendix E.

MMStar questions were presented with the original multiple-choice options, along with Korean translations, in the same order as the dataset.

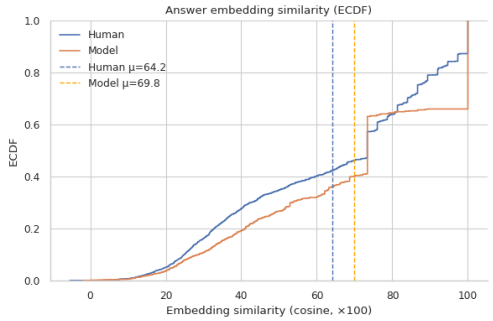
B Effects of Blind Awareness Instruction

After the original prompt and the blank image, the models were instructed as the following: 'Note: No images are provided. For each question, imagine an appropriate image exists and answer based on the most common or universal scenario.'.

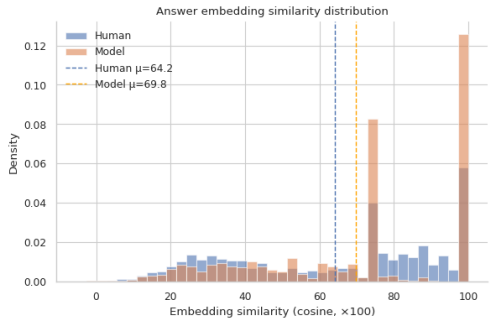
	VQA			MMStar		
Model	GT	Blind	Δ_{Inst}	GT	Blind	Δ_{Inst}
InternVL-3.5 (1B)	80.5	43.6	0.2	43.4	25.7	-1.9
InternVL-3.5 (2B)	82.1	45.5	-0.2	51.7	26.1	0.7
InternVL-3.5 (4B)	83.0	45.0	2.2	64.3	24.5	-2.9
InternVL-3.5 (8B)	86.9	47.1	-1.7	62.7	28.7	-0.8
Qwen-VL-3 (2B)	83.0	45.5	0.7	51.7	23.9	3.3
Qwen-VL-3 (4B)	85.9	45.5	1.9	63.8	21.8	-3.0
Qwen-VL-3 (8B)	89.4	44.3	3.1	66.0	25.2	1.8

Table 2: Accuracy comparison with the presence of additional instruction on VQA and MMStar. **GT** denotes standard visual inference accuracy. $\Delta_{\text{Inst}} = \text{Blind}_{\text{Inst}} - \text{Blind}$ captures the effect of instruction presence.

C Answer Similarity Distributions



(a) ECDF of answer embedding similarity for human and model responses.



(b) Histogram of answer embedding similarity for human and model responses.

Figure 5: Distribution of answer embedding similarity for human and model responses in VQA dataset. ECDF (top) and histogram (bottom) provide complementary views and show consistent rightward shifts for model-generated answers, indicating greater semantic consistency to the ground truth answer in the blind setting.

D Interrater agreement

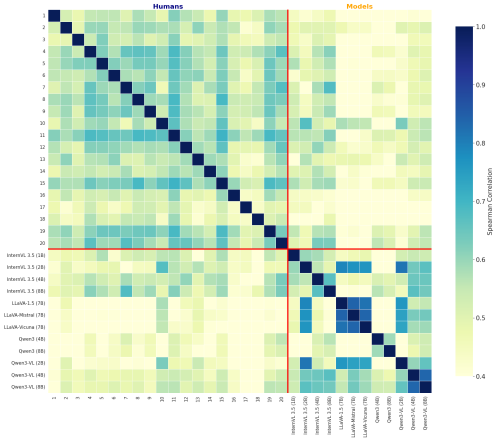


Figure 6: Inter-rater agreement matrix under blind conditions. Each cell reports the pairwise Spearman rank correlation (ρ) between two subjects (human or model), computed over per-question consensus. Higher values indicate stronger rank consistency in question-level behavior, measuring the extent to which the subjects tend to find the same questions easy or difficult in the absence of visual input. Humans and models are separated by a red line.

E Training Details.

Hyperparameters. We apply LoRA to all linear layers with rank 32 and scaling factor 64, while freezing the vision encoder and alignment modules. We optimize the model using AdamW with a learning rate of 2×10^{-5}, cosine learning-rate scheduling with a 5% warmup, and weight decay of 0.1, training for a maximum of 5 epochs. We select the checkpoint with the best validation performance for evaluation. Hyperparameters were selected based on prior literature and preliminary runs, rather than exhaustive optimization.

Answer Aggregation. For training dataset curation, we cluster semantically equivalent answers (e.g., "3" and "three") into canonical forms using using GPT-4o mini (OpenAI, 2025) For each question, we aggregate human responses into a confidence-weighted distribution over canonical answers. Models are then trained on a subset of 15 subject samples and split 20% for validation.

We cluster semantically equivalent human answers using a large language model. The model is prompted to group synonymous responses, select a concise canonical form, and return the result in structured JSON.

The prompt used is shown below.

> **Prompt.** You are an expert VQA data processor. Group semantically similar answers.
>
> **Instructions:** (1) Group semantically identical answers (e.g., "3" and "three"). (2) Choose the most common and concise canonical form. (3) Unique answers form their own group. (4) Return only valid JSON. (5) Ignore unsafe or inappropriate terms.

Confidence Aggregation. Let ϕ denote the mapping from each human answer a_i to its corresponding canonical answer $a_k^* \in \mathcal{A}$. Given a visual question q with image I, we collect N human annotations, where each annotator i provides an answer a_i and a confidence score $c_i \in [0, 1]$.

For each canonical answer a_k^*, we aggregate the confidence scores from all annotations mapped to it:

$$s_k = \sum_{i:\phi(a_i)=a_k^*} c_i \tag{1}$$

The human confidence distribution H over canonical answers is then obtained by normalization:

$$H(a_k^*) = \frac{s_k}{\sum_{j=1}^{K} s_j} \tag{2}$$

Supervised Fine-Tuning Loss. The standard cross-entropy loss on the ground-truth answer sequence:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, q, I) \tag{3}$$

where $y = (y_1, \ldots, y_T)$ is the target answer token sequence.

Distributional Matching Loss. We adopt an answer-level distributional objective that aligns model predictions with human uncertainty over complete answers.

Let M_θ denote the vision-language model with parameters θ, and let $P_\theta(a|q, I)$ denote the model's probability distribution over answers given question q and image I.

For a model M_θ, we define the answer-level predictive distribution

$$M_q(a_j) = \frac{\exp(\ell_j)}{\sum_{k=1}^{K} \exp(\ell_k)}, \tag{4}$$

where $\ell_j = \log P_\theta(a_j \mid q)$ is computed by summing the token-level log probabilities of the answer a_j conditioned on the question q.

We use the Jensen–Shannon divergence (JSD) (Lin, 2002) defined as:

$$\begin{aligned}\mathcal{L}_{\text{dist}}(q) &= \text{JSD}(H_q \,\|\, M_q) \\ &= \frac{1}{2} D_{\text{KL}}(H_q \,\|\, U_q) + \frac{1}{2} D_{\text{KL}}(M_q \,\|\, U_q).\end{aligned} \tag{5}$$

where $U_q = \frac{1}{2}(H_q + M_q)$ is the mixture distribution, and D_{KL} denotes the Kullback–Leibler divergence:

$$D_{\text{KL}}(P \,\|\, Q) = \sum_{a \in \mathcal{A}_q} P(a) \log \frac{P(a)}{Q(a)}. \tag{6}$$

E.1 Effect of supervision under different finetuning strategies on VQA.

Under distributional finetuning, ground-truth supervision yields only marginal improvements in both accuracy and answer similarity across all models. In contrast, standard finetuning method exhibits substantially larger gains from ground-truth supervision, particularly for smaller models.

Model	Strategy	Acc. Δ	Sim. Δ
InternVL 3.5 (8B)	JS	0.37	0.26
	SFT	0.38	0.37
LLaVA–Mistral (7B)	JS	0.22	0.05
	SFT	**0.41**	0.10
Qwen3–VL (4B)	JS	0.49	0.06
	SFT	**1.14**	0.04
Qwen3–VL (8B)	JS	0.09	0.04
	SFT	0.38	0.18

Table 3: Effect of supervision under different finetuning strategies on VQA. We report the difference between ground-truth and blind supervision (GT – Blind) for VQA-trained models.