## References


[1] ShareGPT, 2023.

[2] 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 *NeruIPS*, 2022.

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. 2022.

[6] 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.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICML*, 2021.

[10] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[11] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[12] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL*, 2019.

[13] Google. Bard, 2023.

[14] Google. Gemini, 2023.

[15] 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.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[19] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *NeurIPS*, 2023.

[20] Jennifer Hu and Roger Levy. Prompt-based methods may underestimate large language models' linguistic generalizations. In *EMNLP*, 2023.

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

[22] Heewoo Jun and Alex Nichol. Shap-E: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[25] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[27] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[28] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for GPT-4V (ision), LLaVA-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023.

[29] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

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

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.

[32] 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*, 2023.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017.

[34] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[35] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

[36] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *NAACL*, 2019.

[37] Microsoft. newbing, 2023.

[38] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *ICDAR*, 2019.

[39] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. In *ECCV*, 2022.

[40] OpenAI. GPT-4V(ision) System Card, 2023.

[41] OpenAI. Gpt-4 technical report, 2023.

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[45] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS*, 2021.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

[49] 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.

[50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[51] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020.

[52] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019.

[53] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of MAE pre-pretraining for billion-scale pretraining. In *ICCV*, 2023.

[54] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[55] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *ACL*, 2019.

[56] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

[57] Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models. In *NeurIPS*, 2023.

[58] 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*, 2023.

[59] 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. 2023.

[60] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *NeurIPS*, 2023.

[61] Kirill Vishniakov, Zhiqiang Shen, and Zhuang Liu. Convnet vs transformer, supervised vs clip: Beyond imagenet accuracy, 2024.

[62] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. *arXiv preprint arXiv:2309.16671*, 2023.

[63] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn

of LMMs: Preliminary Explorations with GPT-4V (ision). *arXiv preprint arXiv:2309.17421*, 2023.

[64] 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.

[65] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2022.

[66] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[67] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

[68] 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. *arXiv preprint arXiv:2303.16199*, 2023.

[69] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, 2023.

[70] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *ICLR*, 2021.

[71] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

**Acknowledgements.** We thank Penghao Wu, Muzi Tao, Erik Jones, Michael Psenka, Daniel Yeh, Druv Pai, Chen Sun for helpful discussions and feedback. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. This research is also supported by Intel, Google TRC program, the Google Cloud Research Credits program with the award GCP19980904, and an Amazon Research Award Fall 2023. The authors thank hyperbolic labs for supporting part of the experiments. All experiments and data processing were performed at NYU.

# A. Experiment Details

**Hyperparameters.** In this work, we adopt the same set of hyperparameters as LLaVA [31] and LLaVA-1.5 [30]. We use Vicuna-13b-v1.3 [69] in LLaVA experiments and Vicuna-13b-v1.5 [69] in LLaVA-1.5 experiments. We show the training hyperparameters for LLaVA and LLaVA-1.5 experiments in Table 4. All experiments are conducted using a maximum of 8 Nvidia A100 GPUs.

| Hyperparameter | LLaVA | | LLaVA-1.5 | |
|---|---|---|---|---|
| | Stage 1 | Stage 2 | Stage 1 | Stage 2 |
| batch size | 128 | 128 | 256 | 128 |
| lr | 1e-3 | 2e-5 | 2e-3 | 2e-5 |
| lr schedule decay | cosine | cosine | cosine | cosine |
| lr warmup ratio | 0.03 | 0.03 | 0.03 | 0.03 |
| weight decay | 0 | 0 | 0 | 0 |
| epoch | 1 | 3 | 1 | 1 |
| optimizer | AdamW [33] | | | |
| DeepSpeed stage | 2 | 3 | 2 | 3 |

Table 4. Hyperparameters for MoF training on LLaVA and LLaVA-1.5.

**Pretrain Datasets.** We use the same dataset for both LLaVA and LLaVA-1.5 experiments. For LLaVA experiments, stage 1 uses CC595k [50] and stage 2 uses LLaVA 158k [31] instruction data; For LLaVA-1.5 experiments, stage 1 uses CC595k [50] and stage 2 uses DataMix 665k [1, 15, 21, 23, 24, 31, 34, 35, 38, 49, 51] proposed in Liu et al. [30].

# B. MMVP Benchmark

We provide more details on the MMVP benchmark.

## B.1. Details of evaluating SOTA models

We access GPT-4V through ChatGPT in October and November 2023. We also evaluate Gemini-Pro through Vertex AI API in December 2023. We use the official checkpoints for InstructBLIP [8]. We access mini-GPT4 [71],[1] LLaVA and LLaVA-1.5 [31] through their playgrounds. We test Bard [13] using the official website in September and October 2023. Moreover, we test new-Bing [37] through new-Bing chat creative mode and GPT-4V [40] in September 2023.

## B.2. Questions in MMVP Benchmark

We present more examples in MMVP at the end in Figures 10, 11, 12.

## B.3. Ablation Studies

To further verify that MLLMs make mistakes in MMVP due to their incapable visual grounding instead of hallucination in the language model [20]. We conduct additional ablation experiments on the format and notations of VQA questions and options in MMVP. We choose GPT-4V to do these experiments, as it is currently the best model.

**Swapping options** The first experiment swaps the two options in the MMVP benchmark. For example, we change the question from "Are the butterfly's wings closer to being open or closed? (a) Open (b) Closed" to "Are the butterfly's wings closer to being open or closed? (a) Closed (b) Open".

Empirically, we find that GPT-4V obtains a 40.3% accuracy on the option swapping in our study, as opposed to the original 38.7%. We observe that a few questions are answered differently, while the majority remain the same. This further suggests that the visual incapabilities are in the vision encoder rather than in alignment or the LLMs.

**Changing notations in the options** We conducted an ablation study to assess the impact of altering notations. For example, we changed "(a) Closed (b) Open" to "(1) Closed (2) Open". The results are comparable to the original findings, achieving a performance of 37.3%, closely matching the original 38.7%. The study further suggests that the core challenge in MLLMs is their inherent visual incapability, rather than hallucinations in the language model.

## B.4. Human Study Details

In this study, we ask four participants to volunteer in our study. An example user interface for labeling is shown in Figure 8. We collect their responses and calculate the average score as the human-level performance.

# C. CLIP-MLLM Failure Correlation

**Correlation between CLIP and MLLM models.** We compute the Pearson Correlation between the CLIP model

[1] To circumvent response hallucination in mini-GPT4 we prefix our questions with "Please only choose an option to answer the question below without explanation: "

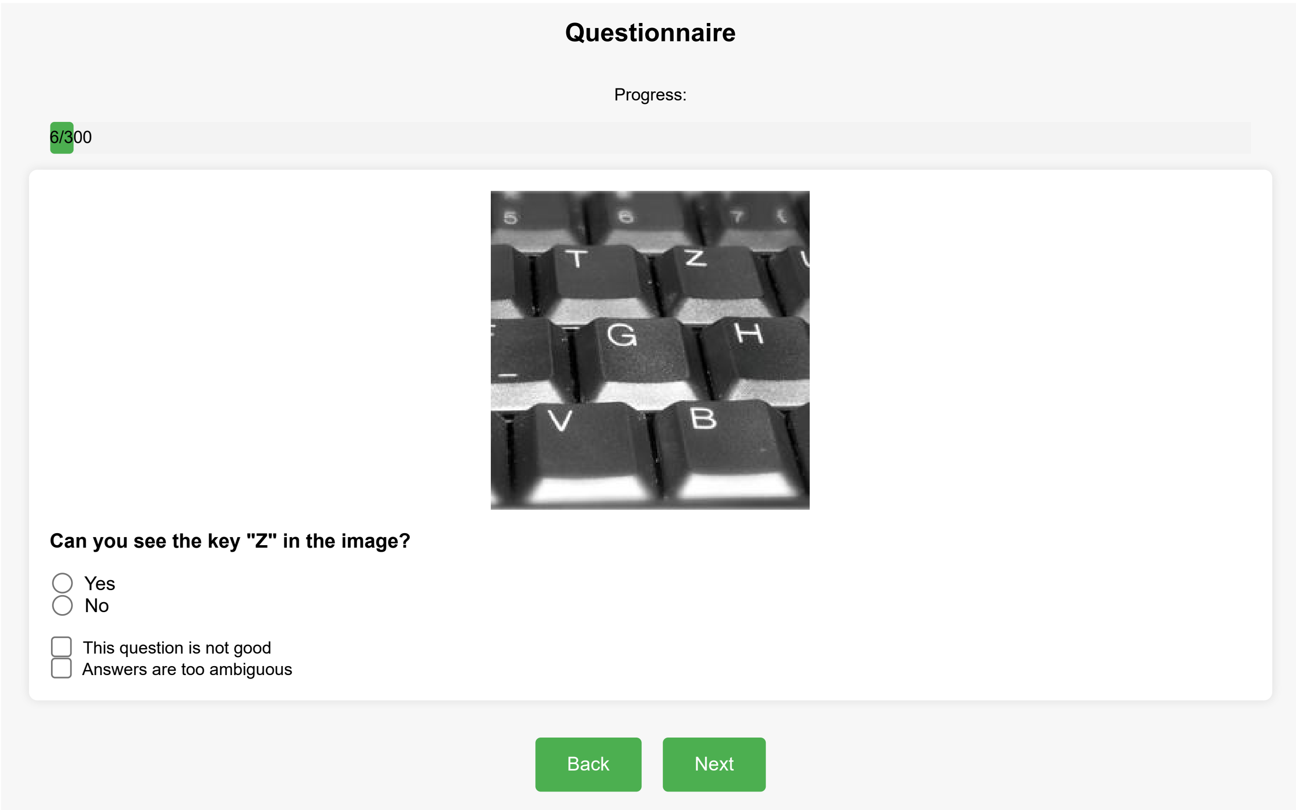


Figure 8. **Example of user study interface.** The questions in the user study are randomly shuffled to avoid any potential bias. Users choose answers for the VQA questions as well as potential concerns for the VQA question.

| | LLaVA-1.5 | InstructBLIP | Bard | Gemini | GPT-4 |
|---|---|---|---|---|---|
| Correlation | 0.87 | 0.71 | 0.79 | 0.72 | 0.31 |

Table 5. Pearson Correlation between the CLIP model and MLLMs. Open-source models that explicitly use CLIP-based models are highlighted in gray.

and MLLMs and show results in Table 5. Notably, both open-source models – LLaVA and InstructBLIP – exhibit remarkably high Pearson Correlation, exceeding 0.7. This finding indicates a strong correlation between the errors made by the CLIP model and those made by MLLMs. Bard also displays a very high correlation. This suggests that some of the most advanced closed-source models are also affected by the visual limitations in the CLIP models.

**Correlation between ImageNet-1k and MMVP performance.** We plot the ImageNet-1k Zero-shot accuracy against MMVP-VLM average performance in Figure 9. For models with ImageNet-1k Zero-shot accuracy below 80, a higher Zero-shot accuracy tends to indicate improved MMVP performance. However, in models with superior ImageNet-1k Zero-shot performance, this trend does not necessarily hold for MMVP-VLM accuracy. This distinction accentuates the value of MMVP-VLM as an evaluation metric, which probes into visual patterns such as orientation – aspects that are pivotal for downstream tasks and go beyond what is captured by ImageNet accuracy alone.

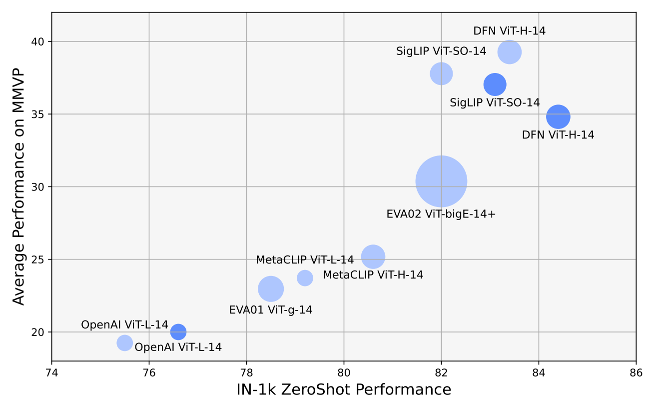


Figure 9. **Correlation between ImageNet-1k Zero-shot and MMVP-VLM average.** The area of each bubble corresponds to the model's number of parameters. A higher ImageNet-1k zero-shot performance does not necessarily imply superior performance in MMVP-VLM.

## D. Visual Patterns for CLIP

Here, we provide the full description of visual patterns that pose challenges to all CLIP-based models.

- **Orientation and Direction**: Questions about the direction something is facing or moving, such as the direction the dog or duck is facing, or the orientation of the school bus.
- **Presence of Specific Features**: Questions that focus on the existence or non-existence of certain elements or features in the image.
- **State and Condition**: Questions that pertain to the state or condition of an object, such as whether a flag is blowing in the wind or if the ground is wet.
- **Quantity and Count**: Questions about the number of objects or features present in the image.
- **Positional and Relational Context**: This aspect refers to the model's ability to understand the position and relationship of objects or elements within an image in relation to each other and their surroundings.
- **Color and Appearance**: Questions regarding the color of certain objects or elements.
- **Structural and Physical Characteristics**: This category involves the model's ability to identify and analyze the physical attributes and structural features of objects in an image.
- **Text**: Questions related to text or symbols present in the image.
- **Viewpoint and Perspective**: Questions concerning the perspective from which the photo was taken.

## E. More Benchmark Results

### E.1. Different vision-only backbones

Here, we conduct extra experiments to study MoF involving MAE [18] or MoCoV3 [17] instead of DINOv2; See Table 6. In Table 6, we observe that with MAE/MoCov3, there is a consistent improvement in visual grounding ability, as shown in the MMVP and POPE benchmarks.

| method | SSL Model | res | #tokens | MMVP | POPE |
|---|---|---|---|---|---|
| LLaVA$^{1.5}$ | None | $336^2$ | 576 | 24.7 | 85.9 |
| LLaVA$^{1.5}$ + I-MoF | MoCov3 | $224^2$ | 512 | 26.7 (+2.0) | 86.1 |
| LLaVA$^{1.5}$ + I-MoF | MAE | $224^2$ | 512 | 27.3 (+2.6) | 86.1 |
| LLaVA$^{1.5}$ + I-MoF | DINOv2 | $224^2$ | 512 | 28.0 (+3.3) | 86.3 |

Table 6. Results of Interleaved MoF with different vision-only SSL model

### E.2. Scaling up to larger resolution

We conduct additional experiments on Interleaved-MoF that further scale up the resolution to 336 and evaluate on more benchmarks. The summarized results in Table 7 reveal that Interleaved-MoF achieves comparable performance on most benchmarks while demonstrating improvements in benchmarks focused on visual grounding. We also observe that MMVP are more sensitive to the model's visual capabilities, underscoring the significance of our benchmark in assessing visual proficiency.

| method | res | #tokens | MMVP | $LLV^{B}$ | $LLV^{W}$ | MMB | $VQA^{T}$ | POPE | $VQA^{V2}$ | MM-V |
|---|---|---|---|---|---|---|---|---|---|---|
| $LLaVA^{1.5}$ | $336^2$ | 576 | 24.7 | **84.7** | 70.7 | **67.7** | **61.3** | 85.9 | **80.0** | **35.4** |
| $LLaVA^{1.5}$ + I-MoF | $224^2$ | 512 | 28.0 | 82.7 | **73.3** | 61.6 | 55.3 | 86.3 | 77.3 | 33.5 |
| $LLaVA^{1.5}$ + I-MoF | $336^2$ | 1152 | **31.3** | 81.8 | **73.3** | 65.4 | 58.7 | **86.7** | 79.3 | 34.6 |

Table 7. **Comparison with LLaVA-1.5 on 6 more benchmarks**. Interleaved-MoF LLaVA-1.5 obtains performance on par with the original method while showing improvements on benchmarks evaluating visual grounding. Benchmark names are abbreviated due to space limits. $LLV^{B}$: LLaVA Benchmark [31]; $LLV^{W}$: LLaVA-In-the-Wild [30]; MMB: MMBench [32]; $VQA^{T}$: TextVQA[52]; POPE: POPE [27]; $VQA^{V2}$: VQA-v2 [15]; MM-V: MM-Vet [64].

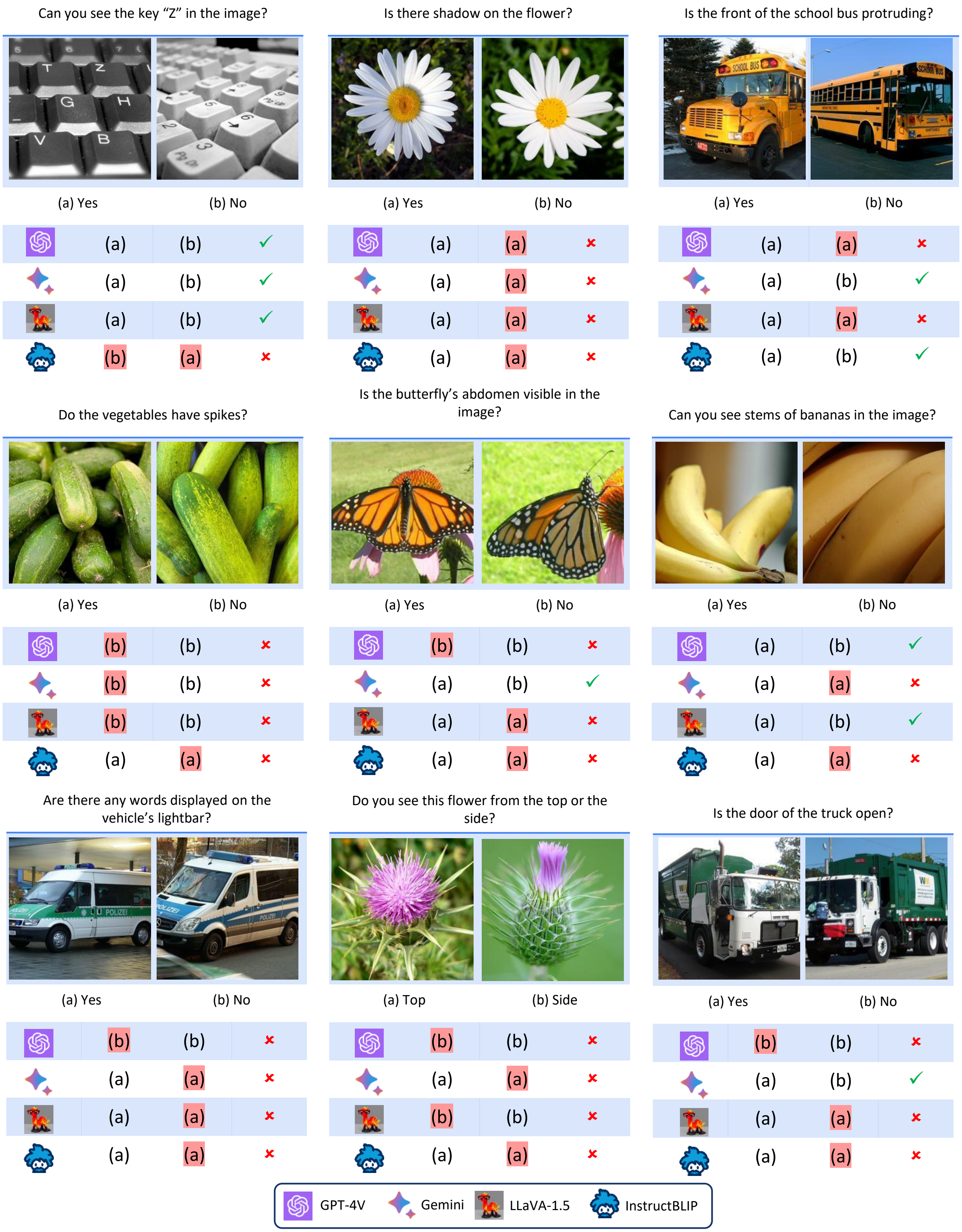


Figure 10. **More examples of questions in the MMVP benchmark (Part I).**

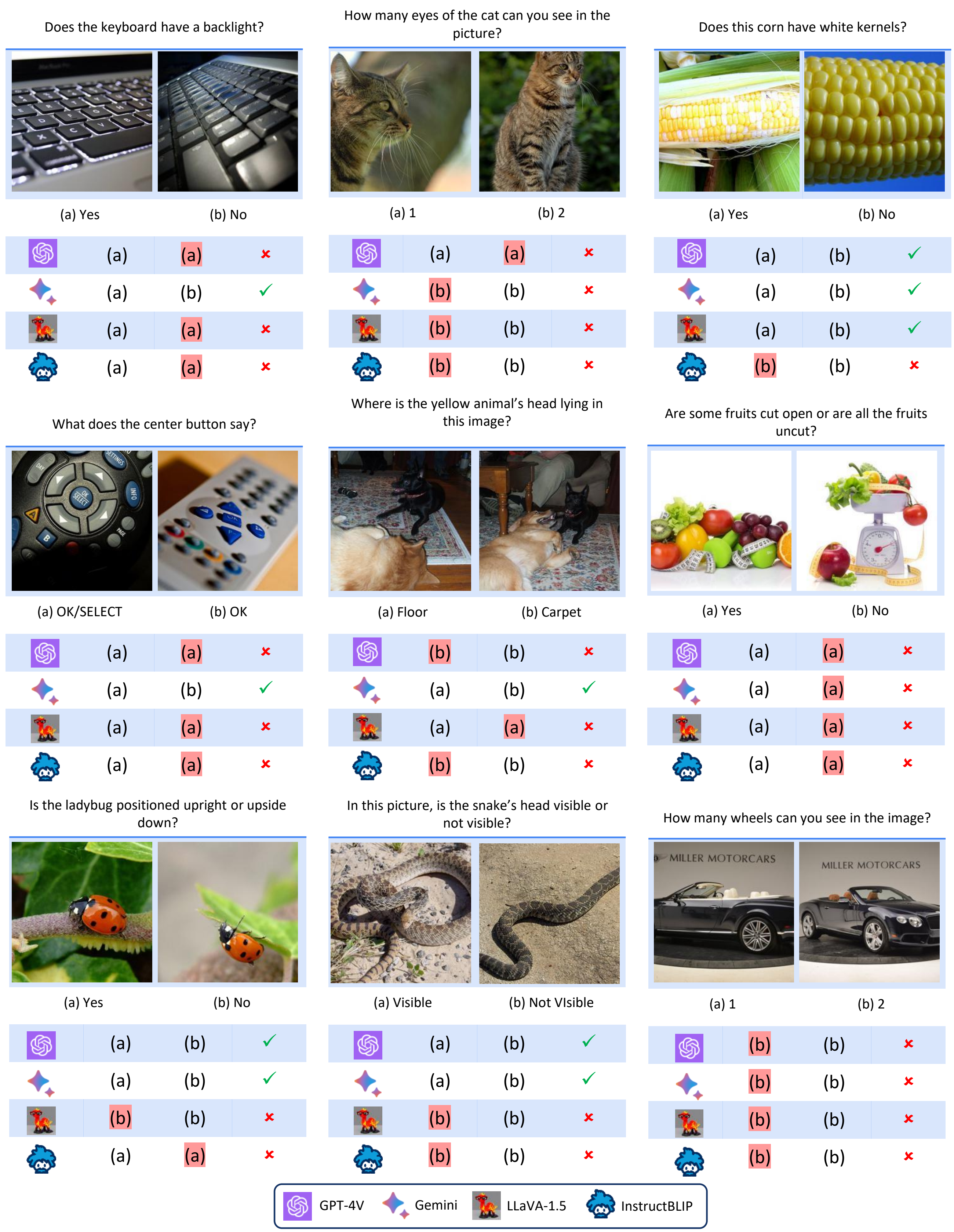


Figure 11. **More examples of questions in the MMVP benchmark (Part II).**

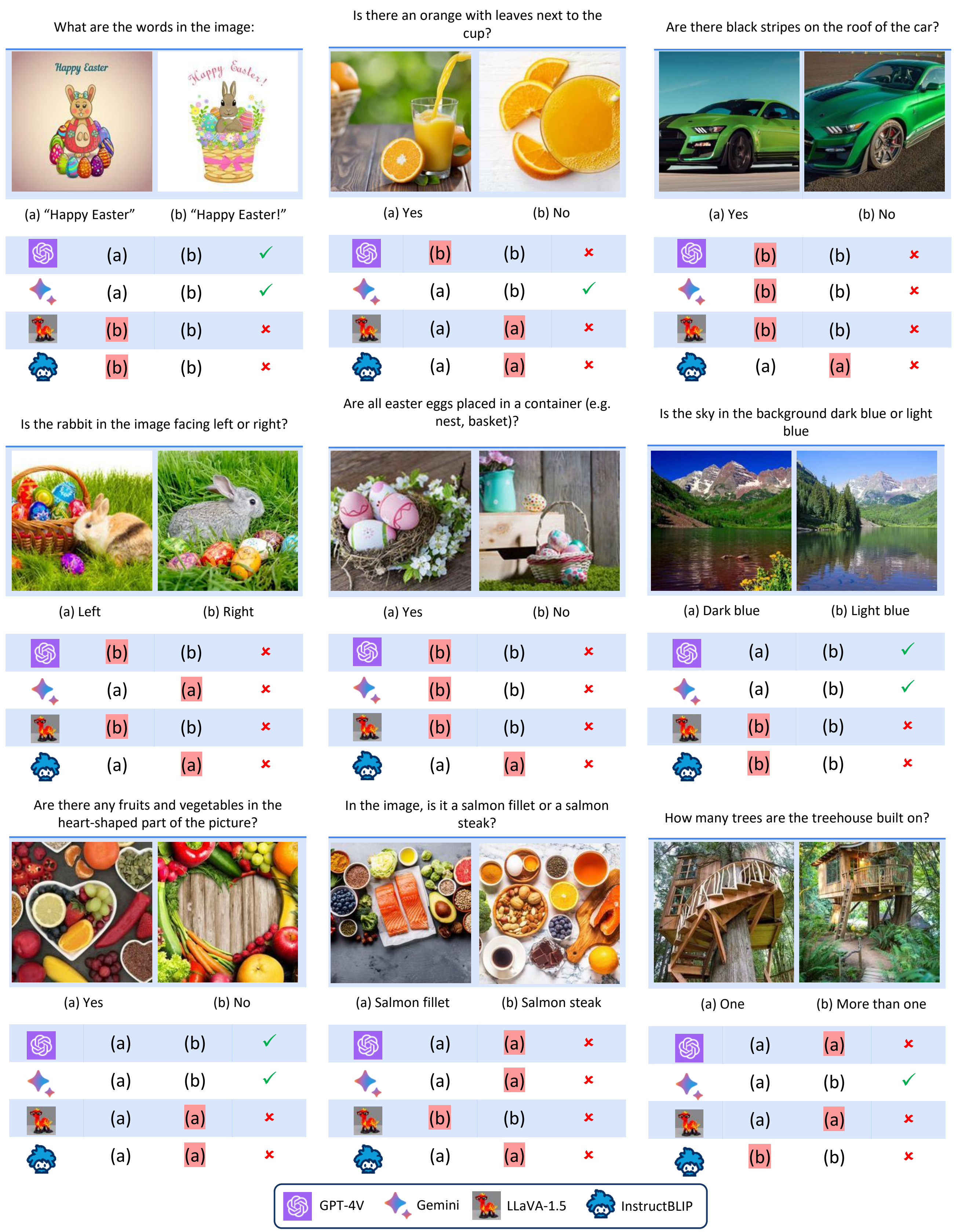


Figure 12. **More examples of questions in the MMVP benchmark (Part III).**