

Seeing the Other Side: Diagnostic Tasks for Viewpoint Reasoning in Vision–Language Models

Makoto Takenaka and Hitomi Yanaka

Mitsubishi Electric, the University of Tokyo, Riken, Tohoku University
Tokyo, Japan

takenaka.makoto@bc.mitsubishielectric.co.jp, hyanaka@is.s.u-tokyo.ac.jp

Abstract

Humans can integrate multiple visual perspectives and infer how an object appears from unseen sides. This study investigates whether Large Vision–Language Models (LVLMs) exhibit a comparable ability for *reference-grounded spatial reasoning*. We propose two diagnostic tasks: **Opposite-Side Reasoning**, which determines whether two images show the same object from opposite viewpoints, and **Viewpoint Identification**, which predicts the viewpoint of a target image using a reference image and its label. An additional condition, **Viewpoint Identification (no-ref)**, removes reference information to reveal cases solvable without it, distinguishing genuine reasoning from bias-driven shortcuts. Our evaluation shows that both open and proprietary LVLMs fall far short of human performance. Even state-of-the-art proprietary LVLMs with relatively high accuracy retain many correct answers when reference information is removed, suggesting that their success often relies on linguistic or dataset-driven priors rather than genuine reference-based reasoning. These findings indicate that current LVLMs have not yet achieved consistent, reference-grounded spatial reasoning. Our datasets in this work will be released on the [Hugging Face Hub](#) to support future research on multimodal viewpoint reasoning and spatial understanding.

Keywords: spatial understanding, Vision–Language Models, viewpoint reasoning

1. Introduction

In recent years, Large Vision–Language Models (LVLMs) have achieved remarkable progress across a wide range of multimodal tasks (Grattafiori et al., 2024; OpenAI et al., 2024; Yin et al., 2024). With their increasing scale and performance, an important challenge has emerged: how to evaluate the *true reasoning capabilities* of these models. Existing benchmarks assess VLMs from various perspectives—such as perception, reasoning, and instruction following (Fu et al., 2024; Yue et al., 2024a,b; Dingjie et al., 2024; Wang et al., 2024; Chen et al., 2024)—yet most of them focus on understanding information that is *directly visible* in the input image.

In contrast, human visual understanding extends beyond what is immediately seen. Humans can infer how an object would appear from unseen sides, grasp its spatial relations, and mentally transform viewpoints. Such abilities represent a form of geometric and relational reasoning that has rarely been examined in existing VLM evaluations. This study aims to investigate the extent to which LVLMs possess this ability to *reason about the unseen sides of objects*.

To this end, we propose two diagnostic tasks to analyze the distinct viewpoint reasoning abilities of LVLMs. The first, **Opposite-Side Reasoning (OSR)**, asks whether two images depict the same object from opposing viewpoints (Figure 1 (left)). The second, **Viewpoint Identification (VI)**, requires predicting the viewpoint of a target im-

age given a reference image and its viewpoint label (Figure 1 (center)). Both tasks are designed not merely to identify visible surface features, but to directly evaluate a model’s ability to *infer how an object appears from unseen directions*.

Additionally, we introduce a supplementary condition for VI tasks, **VI (no-ref)**, in which both the reference image and its textual viewpoint label are removed from the input (Figure 1 (right)). This setting serves as a control condition to detect problems that can be solved correctly *without using reference information*. By comparing VI and VI (no-ref), we can distinguish genuine reasoning based on reference information from incidental correctness arising from internal model biases.

In computer vision, reasoning about the invisible parts of objects has long been recognized as a challenging problem. Previous studies have mainly focused on “amodal completion”, which deals with parts hidden by occlusion (Saleh et al., 2021; Zeng et al., 2020; Mohan and Valada, 2022; Watahiki et al., 2025). In contrast, our work targets viewpoint-dependent invisibility—regions that cannot be seen simply because they fall outside the observer’s perspective. This complementary perspective enables direct evaluation of how well LVLMs can perform *reference-grounded spatial reasoning*.

Our main contributions are summarized as follows:

- We propose two diagnostic tasks, **Opposite-Side Reasoning** and **Viewpoint Identification**, formulated in a VQA task format, to evaluate viewpoint reasoning in LVLMs, and to

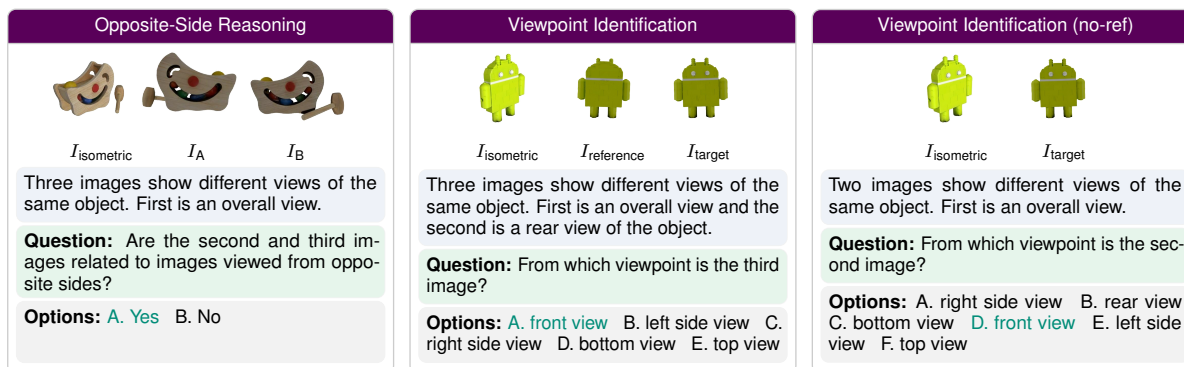


Figure 1: Example from the three tasks. From left to right: Opposite-Side Reasoning, Viewpoint Identification, and Viewpoint Identification (no-ref). Each example consists of two or three images, and the textual prompt is structured into three parts: premise, question, and options. Correct answers are indicated in green.

construct the corresponding datasets, which will be publicly released¹.

- We introduce a controlled condition for VI tasks, **VI (no-ref)**, which isolates reasoning independent of reference information and clarifies the influence of linguistic bias.
- Through evaluations using representative open and state-of-the-art proprietary models, we reveal that current LLMs still struggle to perform visually grounded, reference-consistent reasoning.

2. Proposed Tasks

We propose two diagnostic multi-image reasoning tasks that evaluate distinct yet complementary aspects of viewpoint understanding in LLMs: **Opposite-Side Reasoning (OSR)** (Figure 1 (left)), which probes frame-independent geometric reasoning between object views, and **Viewpoint Identification (VI)** (Figure 1 (center)), which examines whether models can integrate linguistic and visual information in a consistent, spatially grounded manner. Additionally, we introduce a control condition, **VI (no-ref)** (Figure 1 (right)), to identify cases solvable without reference information.

Opposite-Side Reasoning (OSR) The OSR task asks whether two images depict the same object from opposite viewpoints. Each example includes an isometric image $I_{\text{isometric}}$ and two comparison images I_A and I_B ; the model must judge whether I_A and I_B show opposite sides of the same object. This task does not rely on any explicit coordinate frame or linguistic cue; it requires reasoning solely from geometric and structural correspondences between views. OSR therefore measures whether LLMs maintain an internally coherent representation of an object’s 3D geometry that supports

relational reasoning across viewpoints. It is inherently frame-independent, as the correct answer can be inferred without reference to any absolute coordinate system.

Viewpoint Identification (VI) The VI task extends OSR by introducing an explicit linguistic reference, allowing us to assess whether models can jointly use textual and visual information to infer viewpoint relations. Each example includes a isometric image $I_{\text{isometric}}$, a reference image $I_{\text{reference}}$ and a target image I_{target} , typically depicting opposite sides of the same object. The prompt specifies the viewpoint of the reference and asks the model to identify the viewpoint of the target image from five options $\mathcal{V} \setminus \{v_{\text{ref}}\}$, where $v_{\text{ref}} \in \mathcal{V}$ and $\mathcal{V} = \{\text{left side view, right side view, top view, bottom view, front view, rear view}\}$. Although textual information provides an external cue, the task remains frame-independent: It can be solved by understanding the relative correspondence between viewpoints (e.g., right \leftrightarrow left, top \leftrightarrow bottom), rather than by aligning to an absolute coordinate frame. The VI task is thus designed to evaluate whether a model can ground linguistic viewpoint descriptions in visual evidence.

VI (no-ref) The **VI (no-ref)** condition is an auxiliary task designed to complement the VI task. In this task, both the reference image $I_{\text{reference}}$ and the textual prompt describing its viewpoint label v_{ref} are excluded from the VI task. The set of answer options covers the entire viewpoint space \mathcal{V} . This configuration serves as a control condition designed to detect questions that can be answered correctly without using reference information. A high accuracy in VI (no-ref) suggests that the model relies not on reference-based reasoning, but on prior regularities or statistical biases inherent in the dataset. Therefore, the VI (no-ref) score itself is not a direct measure of reasoning ability; its primary purpose is to be compared with the VI results to quantify how

¹<https://huggingface.co/datasets/mtake/Seeing-the-Other-Side>

effectively the model exploits reference information when performing viewpoint reasoning.

Relation and Significance of OSR, VI, and VI (no-ref) OSR and VI serve as complementary probes of spatial reasoning in LVLMs. OSR isolates geometric reasoning that operates without any linguistic or contextual frame, while VI tests whether this visual relation can be coherently aligned with textual descriptions of viewpoint. By introducing VI (no-ref) as a control, we can further disentangle evidence-based reasoning—cases where reference information is genuinely used—from prior-driven correctness, where models answer correctly without relying on the reference.

3. Dataset Construction

We constructed our dataset by rendering images from the Google Scanned Objects (GSO) dataset (GoogleResearch, 2022), which contains 3D models of about 1,000 household objects and is published under the CC BY 4.0 License ².

For the OSR task, we randomly selected 194 objects from the GSO dataset. For each object, we used the isometric image provided in GSO. We rendered two additional opposing-view images in Blender (Blender Online Community, 2025) at a resolution of 512×512 pixels, forming a pair of opposite viewpoints. Each rendered image was assigned one of six viewpoint labels \mathcal{V} based on an egocentric coordinate frame. Positive pairs consisted of true opposite pairs (e.g., left–right), while negative pairs were generated by replacing one image in a positive pair with a non-opposing view (e.g., left–top). In total, the dataset contains 1,164 examples, comprising 582 positive pairs and 582 negative pairs. A uniform textual prompt was used for all examples.

The dataset for the VI task was constructed using the positive pairs from the OSR dataset. For each positive pair (I_A, I_B) , we used I_A as the reference image $I_{\text{reference}}$ and I_B as the target image I_{target} . To account for both directional relations between the two images, the dataset was augmented by including the reversed correspondence (I_B as $I_{\text{reference}}$ and I_A as I_{target}), resulting in a total of 1,164 examples. The textual prompt describes the viewpoint $v_{\text{reference}} \in \mathcal{V}$ of $I_{\text{reference}}$ and presents five answer options $\mathcal{V} \setminus \{v_{\text{reference}}\}$ in random order.

²<https://creativecommons.org/licenses/by/4.0/>

4. Experiments and Results

4.1. Experimental Setting

LVLMs We conducted experiments using three major proprietary LVLMs (GPT-5 (Singh et al., 2025), Claude Sonnet-4 (Anthropic, 2025), and Gemini-2.5 (Gemini Team et al., 2024)) and a diverse set of open LVLMs. For open LVLMs, selection was guided by two criteria: (i) relevance to the current state of the art and (ii) support for multi-image inputs. Since our tasks require reasoning over multiple images simultaneously, we prioritized open LVLMs that natively support multi-image inputs. We included 13 multi-image LVLMs covering three major families (InternVL3_5 (Wang et al., 2025a), Qwen2.5-VL (Bai et al., 2025), and Deepseek_vl2 (Wu et al., 2024)), along with additional LVLMs such as Ldefics3 (Laurençon et al., 2024), Mantis (Jiang et al., 2024), Pixtral (Agrawal et al., 2024), and LLaVA-next (Liu et al., 2024b; Li et al., 2024; Liu et al., 2024a).

For comparison, we also evaluated two single-image LVLMs (LLaVA-v1.5 (Liu et al., 2023) and MiniCPM-Llama3-V (Yao et al., 2024)). Because these models do not natively support multiple-image inputs, we concatenated the input images into a single composite image prior to inference.

Evaluation Protocol Each task is formulated as a Multiple-Choice Question (MCQ), where the LVLM must select the correct answer from several options. To mitigate the known *selection bias* of LVLMs (Atabuzzaman et al., 2025)—such as a tendency to favor certain option symbols (e.g., “A”) or positions—we create three datasets with randomly shuffled option orders for each task, and report the average accuracy across the three runs. For consistent response formatting, each prompt ends with “Select the correct option and enclose your final answer between <answer> and </answer>.” Model outputs are parsed by extracting the text enclosed in these tags and matching either the option symbol or the corresponding label. If the tags are absent, we check model-specific patterns such as “Answer:” or “Answer is.”

We also compare the models’ performance against the random-choice baselines and human evaluation. Human evaluation was conducted by two adult annotators on randomly sampled subsets, consisting of 300 instances for OSR and 150 instances each for VI and VI (no-ref). The annotators achieved average accuracies of 97.2% inter-annotator agreement, measured by Cohen’s κ , was 0.927 for OSR, 0.983 for VI, and 0.567 for VI (no-ref). The moderate agreement observed in VI (no-ref) suggests that this condition introduces greater ambiguity even for human annotators, consistent with the lower accuracy under this setting.

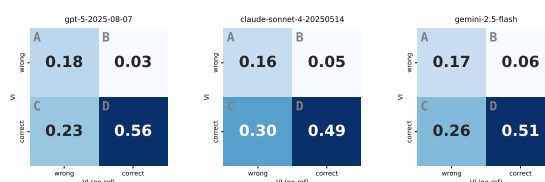
Table 1: Performance comparison on the proposed tasks. We evaluated three paraphrased prompts to assess prompt sensitivity; however, the variations had negligible effect on the overall results.

Model	OSR	VI	VI (no-ref)
Open LVLMS (Multi-Image)			
InternVL3_5-4B	58.3	46.3	29.8
InternVL3_5-8B	51.4	50.7	31.1
InternVL3_5-38B	49.3	59.8	42.3
Qwen2.5-VL-3B-Instruct	49.3	52.1	28.7
Qwen2.5-VL-7B-Instruct	53.2	65.8	39.2
Qwen2.5-VL-32B-Instruct	64.1	54.8	41.4
Mantis-8B-clip-llama3	50.6	44.3	21.4
Pixtral-12B	49.6	52.4	25.7
Idefics3-8B-Llama3	49.4	20.0	16.7
lava_next_interleave_7b	50.0	54.9	28.1
deepseek_vl2_tiny	49.4	45.0	31.6
deepseek_vl2_small	51.1	56.5	43.8
Open LVLMS (Single-Image)			
MiniCPM-Llama3-V-2_5	49.8	40.3	29.4
lava_v1.5_7b	50.0	28.6	18.9
Proprietary LVLMS			
gpt-5-2025-08-07	69.9	79.4	58.7
gpt-5-nano-2025-08-07	50.8	23.4	18.9
claude-sonnet-4-20250514	76.6	79.3	54.4
gemini-2.5-flash	74.7	77.2	57.1
random	50.0	20.0	16.7
human	97.2	99.4	74.7

4.2. Results

The results are shown in Table 1. Across the open LVLMS, performance on the OSR task remained near chance level (50%) for all models except Qwen2.5-VL-32B, indicating limited ability to maintain geometrically consistent multi-view representations. In contrast, proprietary LVLMS achieved substantially higher accuracies, suggesting that large-scale instruction tuning improves opposite-side reasoning. However, a considerable gap remains between these LVLMS and human performance, highlighting that consistent 3D structural understanding across viewpoints is still underdeveloped.

For the VI task, performance varied markedly across LVLMS. Among open LVLMS, InternVL and Deepseek exhibited modest improvements with larger parameter scales, whereas the Qwen family did not show a consistent size-dependent trend. Thus, no clear correlation between model capacity and viewpoint reasoning emerged. Under the VI (no-ref) condition, all LVLMS showed a substantial decrease in accuracy relative to the full VI setting, but proprietary (excluding GPT-5-nano) maintained comparatively high scores. High performance in VI (no-ref) implies an alignment between the model’s internal representation and the egocentric annotation frame used in our dataset. This observation is qualitatively consistent with previous findings that LVLMS exhibit a bias towards egocentric spatial frame (Zhang et al., 2025), suggesting that such bias partially contributes to their viewpoint judgments.



(a) GPT-5 (b) Claude sonnet 4 (c) Gemini 2.5 flash

Figure 2: Confusion matrix of correct and incorrect cases across the two conditions: VI and VI (no-ref) on proprietary LVLMS.

4.3. Open vs. Proprietary LVLMS

A consistent performance gap is observed between open and proprietary LVLMS in all three tasks. On OSR, open models remain 50–53%, while proprietary models reach 70–77%, yielding a margin of more than 20 percentage points. In both VI and VI (no-ref), proprietary systems similarly outperform open models by approximately 15–20 points, indicating that the disparity persists across task settings.

This gap may be influenced by differences in the model scale, architectural design, and the size and diversity of the pretraining data. Larger multimodal pretraining corpora and more advanced fusion mechanisms may support more coherent multi-view representations. However, since the detailed training configurations of the proprietary models are not publicly available, these interpretations remain speculative. Moreover, the lack of a clear scaling trend within open models suggests that parameter size alone may not fully explain the observed differences.

4.4. Analysis of Reference Utilization

Correct predictions in the VI (no-ref) condition indicate that a model may answer correctly without relying on external reference information, such as the reference image or its viewpoint label. Therefore, even if the same instance is also answered correctly in the VI condition, the correctness cannot necessarily be attributed to the use of reference information.

To disentangle the effect of reference information, we compare model predictions between the VI and VI (no-ref) settings and categorize each instance into four regions based on prediction correctness in the two settings: Region A (incorrect in both conditions), Region B (correct only in VI (no-ref)), Region C (correct only in VI), and Region D (correct in both conditions).

The region C represents cases that are correctly answered only when reference information is available, providing a conservative lower bound on the proportion of samples for which the model uses the reference image and its viewpoint label. An example instance in Region C is shown in Appendix A.1.

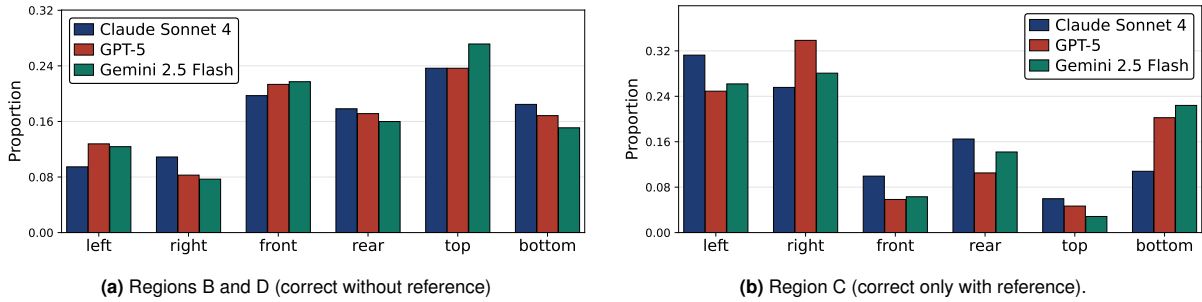


Figure 3: Viewpoint label distributions for proprietary LVLMs; GPT-5, Claude Sonnet 4, and Gemini 2.5 Flash.

Figure 2 shows the confusion matrices for these regions using proprietary models. The results reveal that even state-of-the-art LVLMs exhibit only a limited proportion of Region *C* cases, suggesting that their ability to connect linguistic viewpoint descriptions with visual spatial evidence remains limited.

We further analyze whether viewpoint labels exhibit different levels of dependence on reference information by comparing the distributions of viewpoint labels in Region *C* with those in Regions *B* and *D*. An example instance in Regions *D* is shown in Appendix A.2.

As shown in Figure 3, viewpoint labels exhibit systematic differences in reference dependence. Figure 3a shows that the *top* view label appears more frequently in Regions *B* or *D*, indicating that these cases are often solved correctly without reference information. This may be because many objects have a relatively well-defined vertical orientation, making *top* and *bottom* viewpoints easier to infer even without reference information. In contrast, Figure 3b shows that lateral viewpoints such as *left* and *right* appear more frequently in Region *C*, suggesting that distinguishing *left* from *right* often requires reference information to determine object orientation.

Although the exact proportions vary across models, the qualitative pattern remains consistent: *top* tends to occur more frequently in cases solvable without reference information, whereas *left* and *right* appear more frequently in cases where reference information contributes to correct predictions. This pattern suggests that LVLMs may rely on direct visual cues for some viewpoints while relying on relational reference information for others when performing viewpoint reasoning.

5. Related Work

The proposed task suite is closely related to the concept of *mental rotation* in cognitive psychology (Shepard and Jacqueline, 1971; Surtees et al., 2013). Classical studies have shown that humans can mentally rotate objects in a three-dimensional

space and compare their spatial configurations.

Recently, several benchmarks have been introduced to systematically evaluate the spatial reasoning ability of vision–language models through mental-rotation-related tasks. For example, LogicVista (Xiao et al., 2024) and its curated extension SITE (Wang et al., 2025b) include sub-tasks that estimate the viewpoint of 3D polycube objects, thereby assessing spatial visualization skills based on abstract geometric forms. Meanwhile, BLINK (Fu et al., 2024) introduces a multi-view reasoning task using pairs of real-world images, where models judge whether the camera motion between them is clockwise or counterclockwise.

In contrast, our task suite enables evaluation grounded in the visual statistics of the real world by employing real object images. Moreover, by requiring models to infer the appearance of the opposite side of background-free objects, our tasks assess a more stringent mental rotation that cannot rely on background continuity or motion cues. Furthermore, the comparison between VI and VI (no-ref) establishes a new framework for quantifying how effectively models exploit reference information—namely, their capacity for reference-grounded spatial reasoning.

6. Conclusion

We proposed two diagnostic tasks to assess viewpoint reasoning and evaluated them on open and proprietary LVLMs. Open models showed large gaps from humans, while proprietary LVLMs achieved higher accuracy but often succeeded without reference information, revealing reliance on linguistic or dataset biases. These results indicate that current LVLMs have not yet achieved robust, visually grounded viewpoint reasoning. Our task suite provides a foundation for future research on *reference-based spatial reasoning* in LVLMs.

7. Limitations

This study has several limitations. First, we do not have access to the detailed training configurations of proprietary models, including data sources, model scales, and alignment procedures. As a result, comparisons with open models may not reflect fully equivalent conditions, and the observed performance gaps should be interpreted with caution.

Second, due to computational constraints, our evaluation of open models was limited to relatively small-parameter models, which may favor proprietary systems with larger capacity and broader training corpora.

Third, although human evaluation was conducted with two annotators and inter-annotator agreement was reported, the evaluation was performed on sampled subsets due to budget constraints. While agreement was high for OSR and VI and moderate for VI (no-ref), a larger-scale human evaluation would provide a more comprehensive assessment of task difficulty and ambiguity.

8. Ethics Statement

We confirm that this research was conducted using publicly available datasets and software in an appropriate and responsible manner. The dataset employed in this study is a subset of an existing public dataset and contains no personal or sensitive information. No personal data was collected during the annotation process, and all data were reviewed to ensure that they do not include any content that is offensive or inappropriate. We also ensure that the datasets we plan to release contain no harmful or unethical material, and the likelihood of such content being inadvertently included is considered extremely low.

9. Acknowledgements

We thank the three anonymous reviewers and the meta-reviewer for their valuable comments and suggestions. This work was partially supported by JST BOOST Grant Number JPMJBY24H5, Japan.

10. Bibliographical References

- Pravesh Agrawal et al. 2024. [Pixtral 12b](#).
- Anthropic. 2025. Claude 4 System Card. <https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf>. Accessed: 2026-3-5.
- Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. 2025. [Benchmarking and mitigating mcqa selection bias of large vision-language models](#).
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. [Qwen2.5-VL Technical Report](#). *arXiv e-prints*, page arXiv:2502.13923.
- Blender Online Community. 2025. [Blender - a 3D modelling and rendering package](#). Blender Foundation, Blender Institute, Amsterdam.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. [Are we on the right way for evaluating large vision-language models?](#)
- Song Dingjie, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. [MileBench: Benchmarking MLLMs in Long Context](#). In *First Conference on Language Modeling*.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. [BLINK: Multimodal Large Language Models Can See but Not Perceive](#).
- Gemini Team et al. 2024. [Gemini: A family of highly capable multimodal models](#).
- GoogleResearch. 2022. [Scanned objects by google research](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. [The Llama 3 Herd of Models](#).
- Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhui Chen. 2024. [MANTIS: Interleaved Multi-Image Instruction Tuning](#). *Transactions on Machine Learning Research*, 2024.
- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. [Building and better understanding vision-language models: insights and future directions](#). *ArXiv*, abs/2408.12637.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. [LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models](#).

- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. [LLaVA-NeXT: Improved reasoning, OCR, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Rohit Mohan and Abhinav Valada. 2022. [Amodal panoptic segmentation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21023–21032.
- OpenAI et al. 2024. [GPT-4 Technical Report](#).
- Kaziwa Saleh, Sandor Szenasi, and Zoltan Vamossy. 2021. [Occlusion handling in generic object detection: A review](#). In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMII)*. IEEE.
- Roger N. Shepard and Metzler Jacqueline. 1971. [Mental rotation of three-dimensional objects](#). *Science*, 171(3972):701–703.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. [OpenAI GPT-5 System Card](#).
- Andrew Surtees, Ian Apperly, and Dana Samson. 2013. [Similarities and differences in visual and spatial perspective-taking processes](#). *Cognition*, 129(2):426–438.
- Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. [MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding](#).
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025a. [InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency](#). *arXiv preprint arXiv:2508.18265*.
- Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. 2025b. [SITE: towards Spatial Intelligence Thorough Evaluation](#).
- Amane Watahiki, Tomoki Doi, Taiga Shinozaki, Satoshi Nishida, Takuya Niikawa, Katsunori Miyahara, and Hitomi Yanaka. 2025. [Bridging perception and language: A systematic benchmark for LVLMS’ understanding of amodal completion reports](#). In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*, volume 47, pages 4303–4309.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. [DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding](#).
- Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. [LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts](#).
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. [MiniCPM-V: A GPT-4V Level MLLM on Your Phone](#). *arXiv preprint 2408.01800*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. [A survey on multimodal large language models](#).
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2024a. [MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhua Chen, and Graham Neubig. 2024b.

MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark.

Dan Zeng, Raymond N. J. Veldhuis, and Luuk J. Spreeuwers. 2020. [A survey of face recognition techniques under occlusion](#). *IET Biom.*, 10:581–606.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025. [Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities](#). In *The Thirteenth International Conference on Learning Representations*.

A. Appendices

A.1. Reference-Dependent Left/Right Disambiguation.

Figure 4 illustrates an example belonging to region *C*, where the model answers correctly under the VI setting but fails in the VI (no-ref) condition. The task is to determine the viewpoint of a target image showing a sneaker.

In the VI setting, three images are provided: an overall view, a labeled reference image explicitly identified as the “left-side view”, and the target image. The model correctly observes that the target image presents the opposite profile relative to the labeled left view. By comparing the two side profiles, it is inferred that the target must correspond to the right-side view and produce the correct answer.

In contrast, under the no-ref condition, only the overall view and the target image are given. Although the model accurately describes visual details such as the shoe’s tiger-stripe pattern, side profile, and structural components, it incorrectly predicts “left-side view”. Without an explicit reference anchor, the model lacks a stable coordinate frame to disambiguate lateral orientation.

This example demonstrates that left/right distinctions are intrinsically relational and often require explicit reference grounding, even when perceptual matching remains intact.

A.2. Top View Correct Without Reference.

Figure 5 illustrates an example belongs to region *D*, where the model answers correctly even without the reference image $I_{\text{reference}}$. The task is to determine the viewpoint of a woven basket with a white fabric liner.

Under the VI setting, three images are provided: an isometric view, a reference image explicitly labeled as the bottom view, and the target image. The model identifies that the target image shows

the interior of the basket from above. It observes that the opening of the basket is fully visible, the interior space and white fabric liner can be seen, and the metal handles appear on opposite sides. Based on these geometric and structural cues, the model correctly predicts “top view”.

Importantly, under the VI (no-ref) condition, where $I_{\text{reference}}$ is not provided, the model still produces the correct answer. The same visual cues—interior visibility, opening shape, and handle placement—are sufficient to infer that the viewpoint is from above.

This example illustrates that vertically oriented viewpoints, such as top views, can often be determined without explicit reference grounding. In contrast to lateral distinctions, the geometry of interior visibility provides strong visual evidence that enables stable viewpoint inference.



$I_{\text{isometric}}$

$I_{\text{reference}}$

I_{target}

Viewpoint Identification (with $I_{\text{reference}}$)

Input

Three images show different views of the same object. The first is an overall view, the second is a left side view.

Question: From which viewpoint is the third image?

Options:

- A. right side view
- B. top view
- C. rear view
- D. front view
- E. bottom view

Select the correct option and enclose your final answer between <answer> and </answer>.

Output

I need to analyze the three images to determine the viewpoint of the third image.

Looking at the images:

Image 1: Shows an overall/three-quarter view of an orange and black striped high-top sneaker. I can see the front, side, and top portions of the shoe.

Image 2: Shows the left side view of the same sneaker. I can see the profile of the shoe from the left side, showing the side panel with the striped pattern.

Image 3: Shows the same sneaker from what appears to be the opposite side - the right side view. I can see the profile of the shoe from the right side, showing the right side panel with the same striped pattern, but from the mirror perspective compared to image 2.

Comparing images 2 and 3, they show opposite profiles of the same shoe. Since image 2 is explicitly identified as the left side view, image 3 must be the right side view.

<answer>A</answer> ✓

Viewpoint Identification (no-ref)

Input

Two images show different views of the same object. The first is an overall view. Question: From which viewpoint is the second image?

Options:

- A. rear view
- B. right side view
- C. left side view
- D. bottom view
- E. front view
- F. top view

Select the correct option and enclose your final answer between <answer> and </answer>.

Output

I need to analyze both images to determine the viewpoint of the second image. Looking at the first image, I can see a basketball shoe with orange and black tiger-stripe pattern, viewed from what appears to be a three-quarter front angle showing the medial (inner) side of the shoe. In the second image, I can see the same shoe but from a different angle. The shoe is now shown from the side, and I can see:

- The full profile of the shoe from toe to heel
- The lacing system clearly visible
- The ankle collar and heel area
- The sole/midsole structure
- The tiger stripe pattern continuing along the side

Comparing the two images, the second image shows the shoe from a side perspective. Given the orientation and the way the shoe is positioned, this appears to be the left side view of the shoe (showing the lateral side).

<answer>C. left side view</answer> ✗

Figure 4: Example from region *C* generated by Claude Sonnet 4. The top row shows the input images. In the middle row, the upper panel shows the VI input prompt, and the lower panel shows the corresponding model output. In the bottom row, the upper panel shows the VI (no-ref) input prompt, and the lower panel shows its output. In the no-ref setting, $I_{\text{reference}}$ is not provided. With a labeled reference image, the model correctly infers the “right-side view”; without reference information, it misclassifies the lateral orientation.



Viewpoint Identification (with $I_{\text{reference}}$)

Input

Three images show different views of the same object. The first is an overall view, the second is a bottom view.

Question: From which viewpoint is the third image?

Options:

- A. top view
- B. right side view
- C. front view
- D. left side view
- E. rear view

Select the correct option and enclose your final answer between <answer> and </answer>.

Output

I need to analyze these three images of the same woven basket to determine the viewpoint of the third image.

Looking at the images:

1. First image: Shows an overall/perspective view of a square woven basket with a white fabric liner. I can see the basket from an angled view that shows the front and side, with the white liner visible at the top.
2. Second image: Shows the bottom view of the basket. I can see the flat woven pattern of the basket's base from underneath.
3. Third image: Shows the interior of the basket looking down into it. I can see the white fabric liner draped inside the basket, and I can see down into the basket's interior space. The metal handles are visible on the sides.

Since the third image shows me looking down into the basket's interior from above, this is a top view.

<answer>A</answer> ✓

Viewpoint Identification (no-ref)

Input

Two images show different views of the same object. The first is an overall view.

Question: From which viewpoint is the second image?

Options:

- A. rear view
- B. right side view
- C. left side view
- D. bottom view
- E. front view
- F. top view

Select the correct option and enclose your final answer between <answer> and </answer>.

Output

I need to analyze both images to determine the viewpoint of the second image.

In the first image, I can see a wicker basket with a white fabric liner. The basket appears to be shown from a three-quarter angle view, allowing me to see the front and one side of the basket. The basket has a square/rectangular shape with woven wicker exterior and a white fabric liner that's folded over the rim.

In the second image, I can see what appears to be the same white fabric liner, but now I'm looking down into the basket from above. I can see the interior of the basket with the white liner fabric, and I can see the handles on opposite sides. The view shows the opening of the basket and the interior space, which indicates this is taken from above looking down.

<answer>F. top view</answer> ✓

Figure 5: Example from region D generated by Claude Sonnet 4. The top row shows the input images. In the middle row, the upper panel shows the VI input prompt, and the lower panel shows the corresponding model output. In the bottom row, the upper panel shows the VI (no-ref) input prompt, and the lower panel shows its output. In the no-ref setting, $I_{\text{reference}}$ is not provided. The model correctly identifies the top view in both settings, indicating that explicit reference information is not required.