MirrorQA: Benchmarking Multimodal LLMs on Mirror-Orientation Reasoning

Jingping Liu[1], **Xingchen Peng**[2], **Yan Zhou**[2], **Ziyan Liu**[2], **Jie Zhai**[2*], **Ronghao Chen**[3*], **Huacan Wang**[4], **Xiaofeng Jia**[5*],
[1]Sun Yat-sen University, Zhuhai, China
[2]East China University of Science and Technology, Shanghai, China
[3]Peking University, Beijing, China
[4]University of Chinese Academy of Sciences, Beijing, China
[5]Bowling Green State University, Bowling Green, USA
liujp68@mail.sysu.edu.cn, zeno.social123@gmail.com

Abstract

Multimodal large language models (MLLMs) have achieved remarkable progress in recent years, yet their ability to perform left–right reasoning in mirror contexts—a fundamental element of spatial cognition—remains underexplored. To address this gap, we introduce MirrorQA, a manually constructed benchmark with 5,549 samples, designed to evaluate MLLMs' capability to distinguish left from right from a subject-centered perspective. MirrorQA is built through a three-stage pipeline (annotation, verification, and final review) to ensure high-quality labeling. Comprehensive evaluations on both open- and closed-source MLLMs show that even the best-performing models achieve only 65.40% accuracy, far below the 99.28% accuracy of humans. These results highlight substantial challenges in current MLLMs when reasoning about left and right, and point to promising directions for future research. MirrorQA and its code are publicly available at anonymous link https://github.com/stargazer-zeno/MirrorQA.

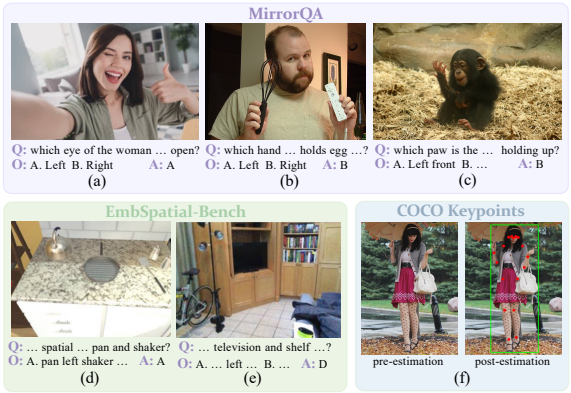

Figure 1: Comparison of sample data from our MirrorQA with EmbSpatial-Bench (Du et al., 2024) and COCO Keypoints (Lin et al., 2014). "Q", "O", and "A" denote the question, options, and answer, respectively.

1 Introduction

Recently, multimodal large language models have demonstrated remarkable performance by integrating strong language understanding with visual perception, achieving impressive results on tasks such as image captioning (Li et al., 2024) and open-ended visual question answering (Kim et al., 2025). However, their limitations become evident when tasks demand finer-grained visual understanding (Liu et al., 2026) and more complex reasoning (Zheng et al., 2023b), highlighting the need for deeper study and targeted improvements in higher-level cognitive abilities.

One striking limitation is their struggle with left–right reasoning in mirror contexts, a basic yet underexplored aspect of visual understanding. For example, when asked "which eye of the woman in the picture is open?" (Fig. 1(a)), models often predict "right eye" instead of the correct answer "left eye", demonstrating their tendency to select the mirror-reversed option. This error stems from their inability to align the observer's coordinate frame with the subject's egocentric frame. Such reasoning is fundamental to human cognition and crucial in safety-critical domains. For instance, in medical imaging, determining whether a fracture is on the patient's left or right side must be made with respect to the patient's body rather than the viewer's perspective—an error in this context could cause severe, even life-threatening, consequences.

To the best of our knowledge, there is currently no benchmark specifically designed for mirror-orientation reasoning. The most relevant efforts lie in two directions: spatial relation reasoning (Liu et al., 2025) and pose estimation (Liu et al., 2024a). While both fields have introduced a variety of high-quality benchmarks, they fall short of comprehensively evaluating MLLMs' abilities to distinguish left from right. Spatial relation reasoning benchmarks, such as EmbSpatial-Bench (Du et al., 2024),

*Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40464–40476

SpatialVLM (Chen et al., 2024a), and VSR (Liu et al., 2023), generally adopt the observer's perspective as the reference frame for spatial judgments. For instance, as shown in Fig. 1(d), the answer "the pan is left of the pepper shaker" defines "left" from the observer's viewpoint. The absence of subject-centered reference frames makes existing benchmarks ill-suited to test whether MLLMs can truly reason about left and right. Pose estimation benchmarks, such as COCO Keypoints (Lin et al., 2014), AIC-HKD (Wu et al., 2017), and AP-10K (Yu et al.), primarily focus on the detection and localization of human or animal keypoints. Their annotation formats—typically joint coordinates—are inherently designed for part identification rather than mirror-orientation reasoning. Consequently, they cannot assess whether models interpret body parts within the subject's own coordinate frame and correctly determine their left–right orientation. For example, although current MLLMs can detect the location of the woman's eyes, they still fail to determine which eye is open (Fig. 1(a)), exposing a key weakness in mirror-orientation reasoning.

Hence, in this paper, we introduce MirrorQA, a multiple-choice benchmark designed to evaluate the mirror-orientation reasoning ability of MLLMs. MirrorQA consists of 5,549 samples based on diverse datasets, such as COCO2017 (Lin et al., 2014) and AP-10K (Yu et al.), with each image paired with a question concerning the orientation of a human or animal body part. The benchmark spans 28 categories of human activities (covering both daily-life and sports scenarios) and 43 animal species, ensuring broad diversity. To guarantee annotation quality, we establish clear guidelines and implement a three-stage human verification process involving annotation, verification, and final review. In addition, symmetric labels (e.g., left/right) are carefully balanced to avoid biases arising from label distribution. For evaluation, we conduct extensive experiments with both open-source models (e.g., Llava (Liu et al., 2023) and Qwen-VL (Bai et al., 2025)) and closed-source models (e.g., ChatGPT-5 and Gemini-2.5-Pro (Comanici et al., 2025)).

In summary, our contributions are threefold:

- We propose MirrorQA, a human-annotated benchmark that addresses the lack of evaluation resources for MLLMs on mirror-orientation (left–right) reasoning.
- The main characteristic of MirrorQA lies in its explicit anchoring of orientation reasoning to the subject's self-centric frame, integrating first-person perspective with spatial reasoning.
- We evaluate both open- and closed-source MLLMs, but even the best models—fine-tuned 7B open-source MLLMs and Gemini-2.5-Pro—reach only 65.40% and 61% accuracy, far below the 99.28% achieved by humans. This highlights the clear challenges MLLMs face in mirror-orientation reasoning.

2 Related Work

Multimodal large language models. In recent years, MLLMs have advanced rapidly, becoming a central focus of research and application. Closed-source models such as OpenAI's GPT series (Achiam et al., 2023) and Google's Gemini series (Comanici et al., 2025) demonstrate strong performance in multimodal tasks, supported by robust technical foundations and large-scale data. Meanwhile, the open-source community has introduced impactful models like Llava (Liu et al., 2023), Qwen-VL2.5 (Bai et al., 2025), and InternVL3 (Chen et al., 2024b), further driving progress in multimodal technologies. These models have shown strong performance in diverse downstream tasks, including visual question answering (Xu et al., 2023), document understanding (Lv et al., 2023), and embodied agents (Shek et al., 2024). Nevertheless, current MLLMs still struggle with mirror-orientation reasoning, especially in identifying body-part orientations relative to the subject. To bridge this gap, we propose a new VQA benchmark designed to evaluate the mirror-orientation reasoning capabilities of MLLMs.

Benchmarks for MLLMs. A variety of benchmarks have been proposed to evaluate MLLMs from different perspectives. The ones relevant to our work fall into two types: pose estimation and spatial relation reasoning. Pose estimation, including both human and animal pose estimation (Zaidi and Wagan, 2024), focuses on localizing anatomical keypoints or body parts from images or videos (Zheng et al., 2023a). Typical datasets such as MPII (Purkrabek and Matas, 2024), AIC-HKD (Wu et al., 2017), AP-10K (Yu et al.), and OpenApePose (Desai et al., 2023) are widely used, but they are designed for keypoint localization and pose recognition rather than for assessing mirror-orientation reasoning. Spatial relation reasoning focuses on a model's ability to infer

object positions, orientations, distances, and relationships (Du et al., 2024). Existing benchmarks include SpatialVOC2K (Belz et al., 2018), SpatialMQA (Liu et al., 2025), and SpatialVLM (Chen et al., 2024a). Even though SpatialMQA includes some perspective transformations, it still emphasizes subject–object relations rather than the orientation of subject's body parts. Hence, we propose MirrorQA, a benchmark for assessing MLLMs' mirror-orientation reasoning abilities.

3 Problem Formulation

In this paper, we formulate mirror-orientation reasoning as a multiple-choice visual question answering task. Given a text question Q and an image I, where Q queries the orientation of a specific body part relative to the subject in the image, the model is required to select the correct answer from k options ($k = 2$ or 4). Each option corresponds to one spatial relation from the predefined set $R = \{$*left*, *right*, *left front*, *right front*, *left rear*, *right rear*$\}$. As shown in Fig. 1(a), given the question "which eye of the woman in the picture is open?", an image, and two options, an ideal model would select "A. left" as the correct answer.

4 MirrorQA Construction

In this section, we describe the construction of MirrorQA in detail, including its image sources, annotation principles, and annotation procedures.

4.1 Image Source

In MirrorQA, we use images of real-world organisms—humans or animals—as the primary subjects, since they possess clear and consistent bilateral structures (e.g., hands, eyes, or ears) that provide an objective basis for left–right orientation. Based on this principle, we collect images from publicly available large-scale datasets. For human subjects, we rely on four datasets: COCO2017 (Lin et al., 2014), UTKFace (Zhang et al., 2017), Leeds Sports Pose (LSP),[1] and 70 Sports.[2] COCO2017 provides diverse everyday scenes and human poses, UTKFace adds over 20,000 face images spanning ages 0–116, and LSP and 70 Sports contribute dynamic sports poses such as gymnastics, fencing, and rowing, further enriching the variety of human non-standard postures. For animal subjects, we use AP-10K (Yu et al.) and Animals-10.[3] AP-10K covers 54 species with high-quality images from the wild, and Animals-10 provides images of common domestic animals, including dogs, cats, and horses.

We further perform a strict quality control on the collected images. Samples that are blurry, too dark, heavily occluded, or lack a clear subject are excluded, ensuring that only images suitable for precise left–right orientation queries remain. In total, the final set contains 20,000 high-quality images: 10,000 from COCO2017, 3,000 from UTKFace, 2,000 from LSP and 70 Sports, 3,000 from AP-10K, and 2,000 from Animals-10.

4.2 Annotation Principles

After collecting the images, we construct one QA pair for each, consisting of a question, candidate options, and a ground-truth answer. To ensure annotation quality, we follow three key principles.

First, each question must require strong visual grounding: the answer should be derivable only from the image itself, not from commonsense or prior knowledge. For example, given an image showing a person's wrist with a watch, the question "On which wrist is the person wearing the watch?" would be invalid, since the model could guess based on the prior knowledge that most people are right-handed and typically wear watches on the left wrist, without actually analyzing the visual evidence. Second, answers must be clear and unambiguous. The image should be sufficiently sharp, the referenced body part easily identifiable, and the question framed to distinctly differentiate left from right, avoiding cases where both could be plausible. Finally, the distribution of answers must be balanced. For each body part across the dataset, the proportion of correct answers referring to the left and right side should remain roughly equal. This prevents the model from developing biased preferences during training and ensures that it genuinely learns left–right body orientation recognition rather than exploiting dataset biases.

4.3 Annotation Procedure

To ensure the benchmark quality, we establish a professional annotation team consisting of three annotators, two checkers, and one final reviewer. In preparation for annotation, all members receive

[1] https://www.kaggle.com/datasets/dkrivosic/leeds-sports-pose-lsp

[2] https://aistudio.baidu.com/projectdetail/3479937?channelType=0&channel=0

[3] https://www.kaggle.com/datasets/alessiocorrado99/animals10

standardized training and conduct a small trial annotation with group discussions to ensure consistent understanding of the task. The benchmark construction proceeds in three stages.

Stage 1: Annotation. Three undergraduate annotators are recruited and evenly assigned 20,000 images. Following the above annotation principles, each annotator manually generates a question, a set of candidate options, and a ground-truth answer for each image. To improve efficiency and reduce workload, we design question templates for different body parts. For instance, the template for the human ear is: "In the picture, which ear does *[object]* have an earring on?"

Stage 2: Verification. Two additional undergraduate checkers examine the outputs from Stage 1, with samples evenly distributed between them. Their primary responsibility is to verify compliance with the three annotation principles and to ensure logical consistency among the question, image, candidate options, and ground-truth answer. When errors are detected, checkers record the issues and return the samples to the original annotators. This iterative process continues until the accuracy of the entire batch exceeds 95%.

Stage 3: Final Review. In the final stage, one of the main authors serves as the final reviewer and performs quality control by randomly sampling 20% of the data that pass the second stage. If any issues are identified, the reviewer records the reasons and returns the problematic samples to the checkers, who then reassign them to the annotators for correction. The process continues until each batch reaches an accuracy rate of at least 98%.

5 MirrorQA Analysis

Benchmark Statistics. As shown in Table 1, MirrorQA contains 5,549 samples, each based on a single image, and they are split into training, development, and test sets in an approximately 7:1:2 ratio. These samples include 3,629 human instances (65.40%) and 1,920 animal instances (34.60%). Each sample is annotated with a label in the format *[subject].[label]*, where *subject* denotes the type (human (H.) or animal (A.)) and *label* specifies the body part orientation. For example, if the answer to "Which hand is the person raising?" is the left hand, the sample is labeled as H.left; if the answer to "Which leg is the horse lifting?" is the left foreleg, it is labeled as A.left front. To mitigate potential bias, we ensure that the left–right orientations of each body part are distributed as evenly as possible. As shown in the lower part of Table 1, MirrorQA achieves a well-balanced label distribution, with the largest difference between symmetric categories (A.left vs. A.right) being only 0.04%.

	Train	Dev	Test	Total	Ratio
MirrorQA	3,880	562	1,107	5,549	100.00
Subjects					
Human	2,539	365	725	3,629	65.40
Animal	1,341	197	382	1,920	34.60
Label Distribution					
H.left	1,270	182	363	1,815	32.71
H.right	1,269	183	362	1,814	32.69
A.left	184	27	52	263	4.74
A.right	182	27	52	261	4.70
A.left front	249	37	71	357	6.43
A.right front	249	36	70	355	6.40
A.left rear	238	35	68	341	6.15
A.right rear	239	35	69	343	6.18

Table 1: Statistics of our MirrorQA. "H." and "A." denotes body parts in human and animal samples.

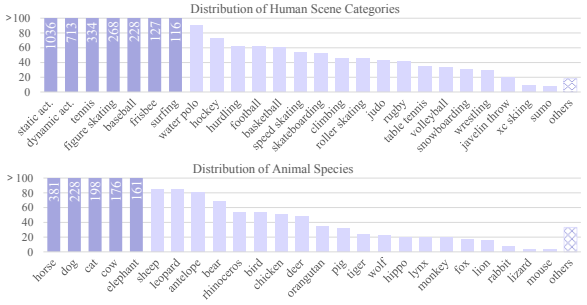

Figure 2: Distribution of human scene categories (top) and animal species (bottom) in MirrorQA.

Image Diversity. MirrorQA covers a diverse range of visual content from both human and animal domains. The human subset includes 28 scene categories, which can be grouped into daily-life and sports activities. Daily-life scenes comprise 1,749 images in total, with 1,036 static activities (e.g., standing, sitting) and 713 dynamic activities (e.g., eating, walking). Sports scenes include 1,880 images across 26 categories, with tennis (334 images), figure skating (268), and baseball (228) being the most frequent. The animal subset contains 1,920 images spanning 43 species, where horses (381 images), dogs (228), and cats (198) are the top three. As shown in Figure 2, for clarity, we visualize the top 25 categories from each subset and merge the remaining less frequent categories into an "others" class: the human "others" class includes 3 cate-

Model	Setting	P	R	F1	Acc	Setting	P	R	F1	Acc
Open-source MLLMs										
InstructBLIP-7B	Base	42.74	47.88	45.16	47.88	LoRA	52.63	49.58	51.06	49.59
Llama-3.2-11B-Vision	Base	43.01	47.49	45.14	47.43	LoRA	51.32	51.53	51.43	51.49
MiniCPM-V-2.6-8B	Base	43.56	46.55	45.01	46.52	LoRA	50.82	51.60	51.21	51.58
Llava-v1.6-7B	Base	46.77	46.64	46.71	46.70	LoRA	66.14	65.33	65.74	**65.40**
DeepSeek-VL-7B	Base	48.25	48.10	48.18	48.06	LoRA	50.13	50.09	50.11	50.05
Phi-3.5-Vision	Base	35.33	45.93	39.94	45.98	LoRA	50.75	50.68	50.72	50.68
Janus-Pro-7B	Base	44.36	44.05	44.21	44.08	LoRA	43.73	48.25	45.88	48.33
Qwen2.5-VL-7B	Base	43.76	45.23	44.49	45.26	LoRA	52.47	51.99	52.23	52.00
InternVL3-8B	Base	43.43	43.64	43.54	43.63	LoRA	50.41	50.21	50.31	50.23
mPLUG-Owl3-7B	Base	38.54	41.91	40.15	41.92	LoRA	58.15	59.38	58.76	59.44
Closed-source MLLMs										
ChatGPT-5	0-shot	48.92	48.92	48.92	49.00	2-shot	50.37	50.42	50.39	50.50
	1-shot	49.88	49.92	49.9	50.00	3-shot	51.76	51.95	51.85	52.00
Gemini-2.5-Pro	0-shot	59.35	59.30	59.33	59.50	2-shot	60.00	59.84	59.92	60.00
	1-shot	59.92	59.92	59.92	**60.00**	3-shot	61.25	60.96	61.10	61.00
Other methods										
Random Choose	-	42.44	42.45	42.45	42.46	-	-	-	-	-
Human	-	99.28	99.28	99.28	99.28	Text-only	40.45	40.49	40.47	40.50

Table 2: Performance (%) of different methods on MirrorQA.

gories (18 images), while the animal "others" class includes 18 categories (33 images). Overall, MirrorQA presents rich and balanced diversity, which supports comprehensive evaluation of MLLMs on mirror-orientation reasoning.

Option Settings. In MirrorQA, answer options are designed in two formats. The first is a binary left/right choice, applied to symmetric human body parts (e.g., eyes, ears, and hands), animals' eyes and ears, and the limbs of birds or other bipedal animals, covering 4,153 samples. The second is a four-way choice—left front, right front, left rear, right rear—applied to animals' limbs, with 1,396 samples. Collectively, these two formats provide a basic coverage of symmetric body-part orientations in real-world organisms, enabling the evaluation of models' mirror-orientation reasoning abilities.

6 Experiments

We conduct a comprehensive evaluation of mainstream MLLMs on our MirrorQA, aiming to assess their performance and identify potential challenges.

6.1 Baselines

To ensure a comprehensive evaluation, we select three categories of baseline methods.

Open-source MLLMs. We select the following baselines: InstructBLIP-7B (Dai et al., 2024), LLaMA-3.2-11B-Vision (Grattafiori et al., 2024), MiniCPM-V-2.6-8B (Yao et al., 2024), Llava-v1.6-7B (Liu et al., 2023), DeepSeek-VL-7B (Lu et al., 2024), Phi-3.5-vision-4B (Abdin et al., 2024), Janus-Pro-7B (Chen et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025), InternVL3-8B (Chen et al., 2024b), and mPLUG-Owl3-7B (Ye et al., 2024). These models are evaluated under two settings: direct inference, which assesses zero-shot performance by generating answers directly from images and questions, and instruction tuning, which fine-tunes them on our training set using the parameter-efficient LoRA method (Hu et al., 2021).

Closed-source MLLMs. We evaluate two SoTA closed-source models, ChatGPT-5 and Gemini-2.5-Pro using in-context learning (ICL) with 0–3 demonstrations. Due to API cost constraints, the evaluation covers 500 samples from MirrorQA.

Other Methods. We further include two baselines: random choice and human evaluation. For the latter, three students (not involved in annotation) answer 500 samples by majority vote, under two conditions: with image and question, or question only.

6.2 Settings and Metrics

The task prompts for both open-source and closed-source MLLMs are provided in Appendix A. All experiments are run on a workstation with four NVIDIA A100-PCIE-40GB GPUs. For evaluation, we use four metrics: precision (P), recall (R), F1,

Model	Setting	H.left	H.right	A.left	A.right	A.L front	A.R front	A.L rear	A.R rear
Open-source MLLMs									
InstructBLIP-7B	LoRA	10.74 (8)	88.95	15.38 (2)	84.62	100.00	4.29 (23)	57.35	33.33
Llama-3.2-11B	LoRA	28.93	77.35	50.00	46.15	4.23 (0)	87.14	19.12 (26)	84.06
MiniCPM-V-2.6-8B	LoRA	39.39	67.13	44.23	61.54	4.23 (1)	84.29	7.35 (1)	91.30
Llava-v1.6-7B	LoRA	60.33	75.14	88.46	76.92	94.37	22.86	23.53	69.57
DeepSeek-VL-7B	LoRA	41.05	61.05	34.62	40.38	40.85	52.86	70.59	44.93
Phi-3.5-Vision	LoRA	62.81	42.27	65.38	23.08	54.93	34.29	60.29	43.48
Janus-Pro-7B	LoRA	86.78	7.18 (39)	94.23	0 (7)	61.97	52.86	22.06	71.01
Qwen2.5-VL-7B	LoRA	74.10	36.19	57.69	21.15	32.39	71.43	38.05	51.94
InternVL3-8B	LoRA	66.39	40.61	76.92	11.54 (17)	52.11	34.29	52.94	36.23
mPLUG-Owl3-7B	LoRA	66.67	56.91	65.38	65.38	53.52	58.57	2.94 (0)	88.41
Closed-source MLLMs									
ChatGPT-5	0-shot	51.56	53.12	30.00	20.00	61.54	38.46	41.67	57.14
	3-shot	51.56	65.62	20.00	10.00	76.92	46.15	33.33	42.86
Gemini-2.5-Pro	0-shot	48.44	64.06	90.00	70.00	92.31	53.85	25.00	64.29
	3-shot	50.00	64.06	80.00	80.00	92.31	69.23	41.67	50.00
Other methods									
Random Choose	-	47.75	49.63	45.51	50.64	23.94	29.05	23.04	20.29
Human	-	99.67	99.00	98.89	98.89	100.00	99.05	99.05	99.05

Table 3: Comparison of results for different body part orientations. Parenthesized numbers indicate the results without fine-tuning and are shown only when the accuracy after fine-tuning falls below 20%. A.L and A.R denote A.left and A.right, respectively.

and accuracy (Acc). All reported results are the average of three runs.

6.3 Main Results

We evaluate all baseline methods on MirrorQA, and the results are reported in Table 2. We observe that 1) Current MLLMs still lag far behind human performance on this task. Even the best-performing model, Llava-v1.6-7B after LoRA fine-tuning, reaches only 65.40% accuracy compared to 99.28% for humans, highlighting the substantial gap between MLLMs and human-level mirror-orientation reasoning. 2) Without fine-tuning, all open-source models perform poorly, with accuracies between 40% and 50%—close to random guessing (42.46%)—indicating that they are almost incapable of solving the task. LoRA-based instruction tuning brings considerable improvements (e.g., Llava-v1.6-7B improves from 46.70% to 65.40%), showing that these models can learn useful mirror-orientation reasoning knowledge from our benchmark, yet their performance remains far from human level. 3) Closed-source models (ChatGPT-5 and Gemini-2.5-Pro) demonstrate stronger in-context learning ability, with accuracy steadily increasing as the number of provided examples grows from 0 to 3 shots, but their best result (61.00% by Gemini-2.5-Pro) is still well below human accuracy. 4) Moreover, in the text-only setting where images are removed, human accuracy drops to 40.45%, close to random guessing, underscoring that MirrorQA critically depends on visual information and thus provides a rigorous benchmark for evaluating multimodal reasoning.

6.4 Detailed Analysis

Performance analysis across different options. As shown in Table 1, each sample in MirrorQA is annotated with a body part orientation, which determines its fixed answer options: H.left and H.right are two-choice questions (options A and B), A.left and A.right are also two-choice (A and B), while A.left front, A.right front, A.left rear, and A.right rear are four-choice (A, B, C, and D). Based on these predefined option groups, we analyze the models' biases across different choices, and the results are presented in Table 3.

We observe that all MLLMs produce highly imbalanced prediction distributions across these orientations, whereas human predictions are balanced. For example, the best-performing model, Llava-v1.6-7B, achieves around 90% accuracy on A.left and A.left front, but only about 20% on A.right front and A.left rear, indicating clear directional bias. To further examine this, we additionally report the corresponding results without

Model	Setting	Vanilla	Circular	ΔAcc	Setting	Vanilla	Circular	ΔAcc
Open source MLLMs								
InstructBLIP-7B	Base	47.88	19.24	-28.64	LoRA	49.59	12.56	-37.03
Llama-3.2-11B-Vision	Base	48.06	27.64	-20.42	LoRA	50.05	28.46	-21.59
MiniCPM-V-2.6-8B	Base	46.52	20.78	-25.74	LoRA	51.58	26.38	-25.20
Llava-v1.6-7B	Base	47.43	5.42	-42.01	LoRA	51.49	6.05	-45.44
DeepSeek-VL-7B	Base	46.70	6.59	-40.11	LoRA	65.40	23.40	-42.00
Phi-3.5-Vision	Base	45.98	11.56	-34.42	LoRA	50.68	24.75	-25.93
Janus-Pro-7B	Base	44.08	10.21	-33.87	LoRA	48.33	5.69	-42.64
Qwen2.5-VL-7B	Base	45.26	11.20	-34.06	LoRA	52.00	15.18	-36.82
InternVL3-8B	Base	43.63	19.78	-23.85	LoRA	50.23	23.76	-26.47
mPLUG-Owl3-7B	Base	41.92	22.76	-19.16	LoRA	59.44	29.54	-29.90
Closed-source MLLMs								
ChatGPT-5	0-shot	49.00	32.00	-17.00	3-shot	52.00	34.00	-18.00
Gemini-2.5-Pro	0-shot	59.50	41.50	-18.00	3-shot	61.00	44.00	-17.00

Table 4: Comparison of Vanilla and Circular evaluations. Vanilla reports single-pass accuracy, Circular reports accuracy with permuted answer options, and ΔAcc is their difference (negative values indicate a drop under Circular).

fine-tuning (shown in parentheses) when the fine-tuned accuracy is below 20%. These results show that most models already exhibit option biases before fine-tuning—for instance, InstructBLIP-7B prefers A and Janus-Pro-7B prefers B in two-choice questions, while MiniCPM-V-2.6-8B tends to choose A and C and mPLUG-Owl3-7B favors C in four-choice questions. After fine-tuning, most models retain their original biases, though exceptions exist. For example, Janus-Pro-7B shows no bias on H.left/H.right before fine-tuning but becomes strongly biased toward A afterward (with only 7.18% accuracy on B). A plausible explanation is that its pre-existing preference for A on A.left/A.right samples is reinforced during fine-tuning and transferred to H.left/H.right questions.

Vanilla evaluation vs. Circular evaluation. To address the limitation of standard multiple-choice evaluation, i.e., positional bias, we adopt a stricter strategy called Circular evaluation (Liu et al., 2024b). This method permutes answer options cyclically and queries the model multiple times; a prediction is correct only if the model selects the right answer in all permutations. A specific example is shown in Appendix B.

Table 4 compares Vanilla and Circular evaluation. Accuracy drops markedly under Circular evaluation across all models, confirming that standard performance is partly inflated by positional bias. Among open-source models without fine-tuning, mPLUG-Owl3-7B shows the smallest decline (-19.16%), while Llava-v1.6-7B suffers the steepest drop, from 47.43% to 5.42% (-42.01%). Even with LoRA fine-tuning, large declines remain: LLaMA-3.2-11B-Vision and MiniCPM-V-2.6-8B show moderate reductions (-21.59% and -25.20%), whereas Llava-v1.6-7B falls by -45.44%. Closed-source MLLMs are also affected, though less severely: ChatGPT-5 drops by 17–18%, and Gemini-2.5-Pro shows similar decreases. Overall, these results demonstrate that positional bias is a pervasive challenge for MLLMs, with closed-source models showing comparatively greater robustness.

Impact of parameter size on model performance. To investigate the impact of parameter scale on MirrorQA, we evaluate two model families: Qwen2.5-VL and InternVL3. As shown in Table 5, both series generally achieve higher accuracy as parameter size increases. For example, Qwen2.5-VL improves from 45.26% (7B) to 58.05% (72B) in the base setting, and further to 59.35% after LoRA fine-tuning. A similar trend is observed in the InternVL3 series, where the 8B and 78B variants obtain 43.63% and 46.79% accuracy in the base setting, rising to 50.23% and 51.76% after LoRA fine-tuning. Despite these improvements, the best-performing models still fall far short of human accuracy (99.28%), underscoring the substantial gap between current MLLMs and human-level mirror-orientation reasoning. Moreover, parameter size is not the sole determinant of performance. Notably, InternVL3-38B underperforms InternVL3-14B in both base and fine-tuned settings, suggesting that factors beyond scale—such as the fine-tuning process and data quality—can exert a significant influ-

Model	Setting	P	R	F1	Acc	Setting	P	R	F1	Acc
Qwen2.5-VL series										
Qwen2.5-VL-7B	Base	43.76	45.23	44.49	45.26	LoRA	52.47	51.99	52.23	52.00
Qwen2.5-VL-32B	Base	52.86	52.33	52.60	52.39	LoRA	57.25	56.14	56.69	56.19
Qwen2.5-VL-72B	Base	58.77	58.06	58.41	58.05	LoRA	59.62	59.35	59.48	59.35
InternVL3 series										
InternVL3-8B	Base	43.43	43.64	43.54	43.63	LoRA	50.41	50.21	50.31	50.23
InternVL3-14B	Base	46.77	46.97	46.87	46.97	LoRA	50.08	50.41	50.25	50.41
InternVL3-38B	Base	45.22	45.81	45.51	45.80	LoRA	47.64	47.98	47.81	47.97
InternVL3-78B	Base	46.61	46.80	46.70	46.79	LoRA	51.82	51.76	51.79	51.76

Table 5: Performance of models with different parameter sizes.

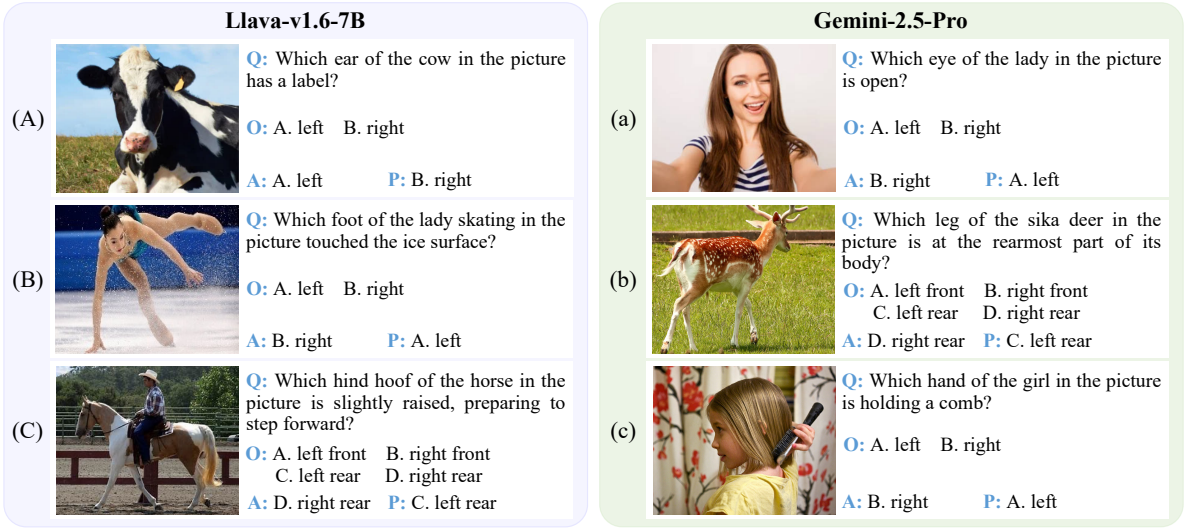

Figure 3: Error examples. Panels (A)–(a), (B)–(b), and (C)–(c) illustrate examples of the RFC, BPM, and OOI error types, respectively. "Q", "O", "A", and "P" denote the question, options, ground truth, and prediction, respectively.

ence on model effectiveness.

6.5 Error Types

To better expose the limitations of current MLLMs in mirror-orientation reasoning and to guide future research, we focus on the two best-performing models from Table 2: the open-source Llava-v1.6-7B and the closed-source Gemini-2.5-Pro. By analyzing their failure cases and conducting manual categorization, we identify three primary error types. The first is Reference Frame Confusion (RFC), where the model confuses the observer's perspective with the subject's own perspective, often resulting in left–right reversal errors (e.g., Fig. 3(a)). The second is Body Part Misidentification (BPM), which reflects difficulties in fine-grained body part recognition. For instance, in Fig. 3(b), the model mistakes the sika deer's left hind leg for the right hind leg—the rearmost one. The third is Obstacle Occlusion Interference (OOI), which arises when key body parts are partially blocked by objects, other limbs, or the surrounding environment, preventing accurate reasoning. An example is shown in Fig. 3(c), where the girl's hand holding the comb is occluded, leading to an incorrect prediction. After manual statistics, RFC is the dominant source of failure, accounting for 51.28% of errors in Llava-v1.6-7B and 55.13% in Gemini-2.5-Pro. The other two error types—BPM and OOI—each account for roughly 33% and 13%, respectively, across both MLLMs. These findings highlight fundamental weaknesses in the mirror-orientation reasoning ability of current MLLMs.

7 Conclusion

We introduce MirrorQA, a benchmark designed to evaluate mirror-orientation reasoning in MLLMs from a subject-centered perspective. It comprises over 5,500 carefully annotated samples encompassing a diverse range of visual content from both human and animal domains. We evaluate a vari-

ety of open- and closed-source MLLMs, and the results show that even advanced models achieve only about 65% accuracy, far below the 99.28% human level. This gap highlights the difficulty of integrating subject-centered reasoning into current MLLMs and underscores that MirrorQA is a valuable benchmark requiring deeper exploration.

Limitations

Although MirrorQA provides a valuable benchmark for evaluating the performance of current MLLMs on mirror-orientation reasoning, it still has several limitations. First, to ensure high data quality, we adopt a rigorous process of manual annotation and human verification. While this approach enhances accuracy, it also limits the overall scale of the benchmark, making it insufficient to support full fine-tuning of MLLMs. Second, although the benchmark strives to capture diverse real-world scenarios—including a broad range of human activities and animal categories—it cannot encompass all possible complexities. In especially challenging cases, such as extreme poses, heavy occlusion, or uncommon camera angles, model robustness remains an open issue that requires further exploration and improvement.

Ethical Statement

Our MirrorQA benchmark is constructed based on COCO2017, UTKFace, LSP, 70 Sports, AP-10K, and Animals-10, all of which allow redistribution and re-annotation as long as the original works are properly cited. Following these requirements, we release MirrorQA under the CC BY 4.0 license. To further ensure responsible use, we have carefully examined the benchmark to confirm that it does not contain harmful content such as gender bias, racial discrimination, or other inappropriate material.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (No. 62306112) and Guangdong Basic and Applied Basic Research Foundation (No. 2026A1515010253).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Anja Belz, Adrian Muscat, Pierre Anguill, Mouhamadou Sow, Gaétan Vincent, and Yassine Zinessabah. 2018. Spatialvoc2k: A multilingual dataset of images with annotations and features for spatial relations between objects. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 140–145.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Nisarg Desai, Praneet Bala, Rebecca Richardson, Jessica Raper, Jan Zimmermann, and Benjamin Hayden. 2023. Openapepose, a database of annotated ape photographs for pose estimation. *elife*, 12:RP86873.

Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355.

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. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Computing Surveys*, 57(10):1–35.

Wenbo Li, Guohao Li, Zhibin Lan, Xue Xu, Wanru Zhuang, Jiachen Liu, Xinyan Xiao, and Jinsong Su. 2024. Empowering backbone models for visual text generation with input granularity control and glyph-aware training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8001–8014.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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

Jian Liu, Wei Sun, Hui Yang, Zhiwen Zeng, Chongpei Liu, Jin Zheng, Xingyu Liu, Hossein Rahmani, Nicu Sebe, and Ajmal Mian. 2024a. Deep learning-based object pose estimation: A comprehensive survey. *arXiv preprint arXiv:2405.07801*.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. 2025. Can multimodal large language models understand spatial relations? *arXiv preprint arXiv:2505.19015*.

Jingping Liu, Xueyan Wu, Hanxuan Chen, Ziyan Liu, Zhangquan Chen, Ronghao Chen, and Huacan Wang. 2026. Easy for children, hard for ai: The limits of multimodal llms in early childhood learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32078–32086.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, et al. 2023. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*.

Miroslav Purkrabek and Jiri Matas. 2024. Improving 2d human pose estimation in rare camera views with synthetic data. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE.

Chak Lam Shek, Xiyang Wu, Wesley A Suttle, Carl Busart, Erin Zaroukian, Dinesh Manocha, Pratap Tokekar, and Amrit Singh Bedi. 2024. Lancar: Leveraging language for context-aware robot locomotion in unstructured environments. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9612–9619. IEEE.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. 2017. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Syed Adnan Ali Zaidi and Asim Imdad Wagan. 2024. Deep learning based approaches for animal pose estimation (ape): A survey. In *2024 International Conference on Emerging Trends in Smart Technologies (ICETST)*, pages 1–6. IEEE.

Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023a. Deep learning-based human pose estimation: A survey. *ACM computing surveys*, 56(1):1–37.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023b. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

A Instruction Settings of MLLMs

The instruction settings for both open-source and closed-source MLLMs are presented in Table 6.

Models	Task instructions
Closed-source MLLMs (zero-shot) and Open-source MLLMs	You are currently a senior expert in visual reasoning. \n Given an Image, a Question, and Options, your task is to choose the correct answer. \n Note that you only need to choose one option from all options without explaining any reason. \n Input: Image: <image>\n Question: {question}, Options: {options}. \n Output: *Note that for InstructBlip, the prompt excludes the image component.*
Closed-source MLLMs (few-shot)	You are currently a senior expert in visual reasoning. \n Given an Image, a Question, and Options, your task is to choose the correct answer. \n Note that you only need to choose one option from all options without explaining any reason. \n Given the following 3 examples to learn the visual reasoning task: \n Example1: \n Image: <image>\n Question: Which eye of the girl is closed?, Options: A. Left; B. Right. \n Output: A. Left \n Example2: ... \n Example3: ... \n Input: Image: <image>\n Question: {question}, Options: {options}. \n Output:

Table 6: Task instructions for MLLMs.

B Circular Evaluation

In CircularEval, each question is tested multiple times through cyclic shifts of the answer options, and the VLM must succeed in all rounds to be correct. As shown in Figure 4, the VLM fails in round 3, so the final answer is judged incorrect.

C Full-parameter Fine-tuning

We additionally evaluate full-parameter fine-tuning on the best-performing model (LLaVA-v1.6-7B). As shown in Table 7, full-parameter fine-tuning provides limited gains over the base model and performs worse than LoRA. We attribute this to the relatively small scale of our dataset, where full-parameter updates are more prone to overfitting, while LoRA remains more stable.

D CoT Prompting

We evaluate the effect of Chain-of-Thought (CoT) prompting on both open- and closed-source models, including LLaVA-v1.6-7B and Gemini 2.5 Pro. The CoT strategy decomposes the task into three stages: role positioning, body-part identification, and decision-making. As shown in Table 8, CoT brings negligible improvement before fine-tuning and leads to a performance drop after LoRA fine-tuning. This suggests that the task involves limited reasoning requirements, and explicit CoT prompting may not be beneficial in this setting.

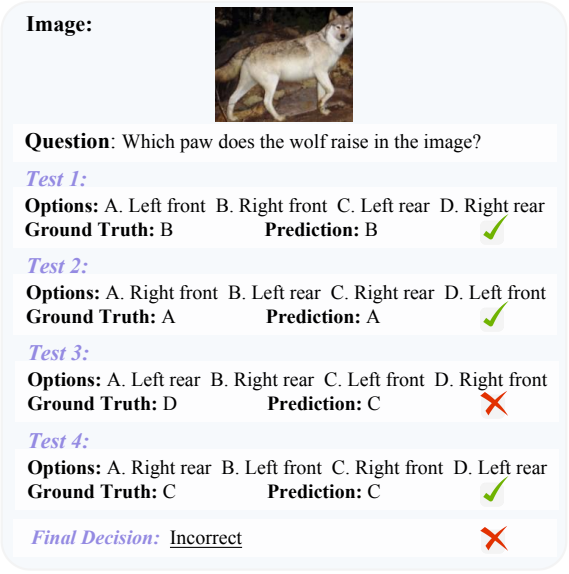

Figure 4: Circular Evaluation.

Model	Setting	P	R	F1	Acc
Llava-v1.6	Base	46.77	46.64	46.71	46.70
	LoRA	66.14	65.33	65.74	65.40
	Full	50.23	50.28	50.25	50.32

Table 7: Full-parameter fine-tuning on our MirrorQA.

Model	Setting	P	R	F1	Acc
Llava-v1.6	Base	46.77	46.64	46.71	46.70
	Base+CoT	48.77	46.91	47.82	46.97
	LoRA	66.14	65.33	65.74	65.40
	LoRA+CoT	59.69	54.97	57.24	55.01
Gemini2.5	Base	59.35	59.30	59.33	59.50
	Base+CoT	59.16	59.00	59.08	59.00

Table 8: Performance comparison with and without CoT prompting on LLaVA-v1.6-7B and Gemini 2.5 Pro.

E Comparison with Spatially Enhanced Models

We further compare our approach with a spatially enhanced MLLM, SpaceLLaVA-7B, which is trained on spatial data. As shown in Table 9,

Model	Setting	P	R	F1	Acc
Llava-v1.6	Base	46.77	46.64	46.71	46.70
	LoRA	66.14	65.33	65.74	65.40
SpaceLLaVA	Base	51.82	17.56	26.23	17.62
	LoRA	49.66	48.76	49.20	48.78

Table 9: Comparison with a spatially enhanced model before and after fine-tuning.

SpaceLLaVA-7B performs substantially worse than the general-purpose LLaVA-v1.6-7B, both before and after fine-tuning. This indicates that spatial data augmentation alone does not effectively address mirror-orientation reasoning in our task.