Can Multimodal Large Language Models Understand Spatial Relations?

Jingping Liu$^{\diamond*}$, Ziyan Liu$^{\diamond}$, Zhedong Cen$^{\diamond}$, Yan Zhou$^{\diamond}$, Yinan Zou$^{\diamond}$,
Weiyan Zhang$^{\diamond*}$, Haiyun Jiang$^{\spadesuit}$, Tong Ruan$^{\diamond}$

$^{\diamond}$School of Information Science and Engineering, East China University
of Science and Technology, Shanghai, China
$^{\spadesuit}$School of Computer Science, Fudan University, Shanghai, China
{jingpingliu,weiyanzhang}@ecust.edu.cn, y30241069@mail.ecust.edu.cn

Abstract

Spatial relation reasoning is a crucial task for multimodal large language models (MLLMs) to understand the objective world. However, current benchmarks have issues like relying on bounding boxes, ignoring perspective substitutions, or allowing questions to be answered using only the model's prior knowledge without image understanding. To address these issues, we introduce SpatialMQA, a human-annotated spatial relation reasoning benchmark based on COCO2017, which enables MLLMs to focus more on understanding images in the objective world. To ensure data quality, we design a well-tailored annotation procedure, resulting in SpatialMQA consisting of 5,392 samples. Based on this benchmark, a series of closed- and open-source MLLMs are implemented and the results indicate that the current state-of-the-art MLLM achieves only 48.14% accuracy, far below the human-level accuracy of 98.40%. Extensive experimental analyses are also conducted, suggesting the future research directions. The benchmark and codes are available at https://github.com/ziyan-xiaoyu/SpatialMQA.git.

1 Introduction

Multimodal large language models have become increasingly significant in AI due to their ability to process and integrate data from multiple sources such as text and images. Although MLLMs excel in tasks like image recognition (Guo et al., 2023) and classification (Wang et al., 2023), they still face challenges with more complex tasks, such as multimodal understanding and reasoning (Zheng et al., 2023), highlighting the need for further exploration and enhancement of their capabilities.

A critical aspect of evaluating MLLMs is their ability to understand spatial relations, which involves inferring the spatial relations between en-
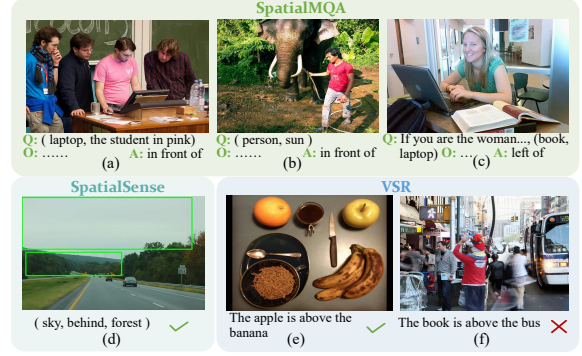

Figure 1: Samples from spatial relation reasoning benchmarks. "Q", "O", and "A" in our SpatialMQA denote the question, options, and answer. In SpatialSense (Yang et al., 2019) and VSR (Liu et al., 2023a), questions are binary classification, with image and text inputs, and true/false outputs.

tities in a given scene (Liu et al., 2023a). For instance, in Figure 1(a), given the subject "laptop" and the object "the student in pink", the model needs to infer that the spatial relation between them is "in front of". This task is important because understanding spatial relations in the objective world is a fundamental human ability essential for daily life (Proulx et al., 2016; Hawes and Ansari, 2020). For instance, to fully the scene of four students discussing at the podium in Figure 1(a), it is necessary to identify the entities (laptop, podium, the student in pink) and their spatial relations (laptop on podium, laptop in front of the student in pink).

Several benchmarks exist for spatial relation reasoning, yet they remain insufficient for fully evaluating MLLMs' ability to understand spatial relations. These benchmarks can be categorized based on whether they use bounding boxes (bboxes) to enclose subjects and objects. First, benchmarks with bbox annotations, such as SpatialVOC2K (Belz et al., 2018), Rel3D (Goyal et al., 2020), and SpatialSense+ (Wen et al., 2024), face two main challenges. On one hand, the subject or object in the question may not be explicitly visible in images,

---
*Corresponding authors.

Table 1: Overview of spatial relation reasoning benchmarks. "Q. Type", "Rel.", "Type", "Obj. W", "Per. sub.", "Kn.", and "MQA" stand for "question type", "relations", "types of subject or object", "objective world", "perspective substitution" "knowledge", and "multiple-choice QA", respectively. "Objective world" indicates whether the benchmark's annotations use the objective world as the reference system. "Perspective substitution" means whether questions involve perspective (first- or third-person). "Knowledge" indicates whether questions in the benchmark can be answered solely with models' prior knowledge, without images.

Benchmark	Q. Type	# Rel.	# Type	w/o bbox	Obj. W	Per. sub.	w/o Kn.	Size
SpatialVOC2K (Belz et al., 2018)	Cloze	17	20	×	×	×	×	2,026
SpatialSense (Yang et al., 2019)	T or F	9	-	×	×	First	×	17,498
Rel3D (Goyal et al., 2020)	T or F	30	67	×	×	×	-	27,336
SpatialSense+ (Wen et al., 2024)	T or F	9	-	×	×	×	✓	7,254
SpatialRGPT (Cheng et al., 2024)	OpenQA	12	88	×	✓	×	✓	1,406
EmbSpatial (Du et al., 2024)	MQA	6	294	✓	×	×	×	3640
VSR (Liu et al., 2023a)	T or F	66	32	✓	Partly	First	×	10,972
SpatialVLM (Chen et al., 2024)	OpenQA	-	-	✓	Partly	×	×	546
SpatialMQA (ours)	MQA	6	128	✓	✓	First/Third	✓	5,392

making it impossible to use bboxes (Liu et al., 2023a). As illustrated in Figure 1(b), the subject "sun" cannot be framed with a bbox in the question "Where is the sun located relative to the man?". On the other hand, some spatial relations in these benchmarks are not grounded in the objective world, leading to a gap between machine and human cognition. For instance, in Figure 1(d), the sky is objectively above the forest, but SpatialSense marks it as behind the forest. Secondly, benchmarks without labeled bboxes, like EmbSpatial (Du et al., 2024), VSR (Liu et al., 2023a), and SpatialVLM (Chen et al., 2024), also face two main issues. One major issue is that they often ignore perspective substitution (first- and third-person). Even when included, it is only a small part. For instance, in VSR, only 6% of the benchmark uses a first-person perspective. This limits the model's ability to understand spatial relations from different perspectives, which is important for complex, dynamic scenarios like autonomous driving (Gao et al., 2024). Another issue is that some questions in these benchmarks can be answered correctly without images, relying only on the model's prior knowledge. As shown in Figure 1(f), the question "the book is above the bus" can often be answered "No" based on commonsense, without needing to analyze the image. This prevents a proper evaluation of MLLMs' image understanding abilities.

Hence, in this paper, we introduce SpatialMQA, a new benchmark in a **m**ultiple-choice **q**uestion & **a**nswer format, designed to fully evaluate the ability of MLLMs in multimodal **spatial** relation reasoning. The benchmark includes 5,392 samples based on COCO2017 (Lin et al., 2014), covering 128 subject and object types, without the use

of bboxes. To address the limitations of existing benchmarks, we establish clear annotation guidelines for SpatialMQA, incorporating questions that involve perspective substitution based on the objective world as a reference system, while avoiding questions that can be answered solely through the models' prior knowledge without images. In addition, we design a three-round annotation procedure for quality control. To assess the spatial relation reasoning capabilities of MLLMs, we conduct comprehensive experiments using closed-source models such as GPT-4o (Achiam et al., 2023) and Gemini-1.5-flash (Team et al., 2023), as well as open-source models like LLaVA (Liu et al., 2024) and SpaceLLaVA (Chen et al., 2024).

In summary, our contributions include:

- We introduce a new manually annotated high-quality benchmark for multimodal spatial relation reasoning without bboxes.

- The main characteristic of SpatialMQA is that the questions involve perspective substitutions using the objective world as a reference. Also, the questions cannot be answered using only the model's prior knowledge without images.

- We evaluate both open- and closed-source MLLMs on SpatialMQA, indicating that state-of-the-art (SoTA) methods like GPT-4o and instruction-tuned SpaceLLaVA achieve accuracies of 40.20% and 48.14%, respectively, far below the human accuracy of 98.40%. We further provide detailed analyses and suggest future research directions.

Table 2: The definition of the spatial coordinate system (SCS) and its six spatial relations. The coordinates for the subject are specified as (x_s, y_s, z_s) and for the object as (x_o, y_o, z_o).

Terms	Definition
SCS	The spatial coordinate system is established based on the objective world, with gravity pointing downward and the observer as the origin. The X-axis spans the observer's left (negative) to right (positive), the Y-axis from back (negative) to front (positive), and the Z-axis from down (negative) to up (positive).
left of right of	The subject is to the left of the object when $x_s < x_o$. The subject is to the right of the object when $x_s > x_o$.
in front of behind	The subject is in front of the object when 1) $y_s \cdot y_o > 0$, $|y_s| - |y_o| < 0$, or 2) $y_s \cdot y_o < 0$, $y_s > 0 > y_o$. The subject is behind the object when 1) $y_s \cdot y_o > 0$, $|y_s| - |y_o| > 0$, or 2) $y_s \cdot y_o < 0$, $y_s < 0 < y_o$.
on/above below	The subject is on/above the object when $z_s > z_o$. The subject is below the object when $z_s < z_o$.

2   Problem Formulation

In this paper, we consider the spatial relation reasoning task as a multiple-choice question-answering problem. Given a text question Q and an image I, where Q asks about the spatial relation between two target entities, the task requires the model to select the correct answer from k ($k = 2, ..., 6$) options. Each option corresponds to a spatial relation from the pre-defined set $R = \{$*left of*, *right of*, *in front of*, *behind*, *on/above*, *below*$\}$, with their definitions provided in Table 2. For instance, in Figure 1(a), given the question "Where is the laptop located relative to the student in pink?", an image, and six options, an ideal model would select "in front of" as the correct answer.

3   SpatialMQA Construction

In this section, we detail the construction of the SpatialMQA benchmark, including image source, annotation guidelines, and annotation procedures.

3.1   Image Source

In this study, we choose COCO2017 (Lin et al., 2014) as our image source due to its notable advantages: 1) Extensive collection. COCO2017 contains over 160,000 images, providing a broad selection for identifying high-quality images to analyze spatial relations. 2) Diverse types: The dataset encompasses 80 entity types, covering a wide range of entities in the objective world, such as people, animals, cars, and food. 3) Multi-entity scenarios: The images in COCO2017 often involve multiple entities, making it easier to select two appropriate entities to determine their spatial relation. From this dataset, we select 30,000 high-quality images to annotate two entities and their spatial relation.

3.2   Annotation Guidelines

Based on the collected images, we label each one with a question, options, and the correct answer. To assist annotators in creating high-quality samples, we provide annotation guidelines, including question types and important precautions.

Question types. Based on the observer's perspective, we divide the question types into two types: out-of-image and in-image perspectives. In the first type, the observer exists outside the image and we manually pre-define several question templates like "Where is the subject located relative to the object?". See Appendix A for more templates. In the second type, the observer's perspective is within the image and can be further divided into two types. The first type uses the object's perspective as the observer's perspective (also denoted as the first-person perspective). Question templates for this type address the spatial relation between the subject and the observer (the object), such as "If you are *[object]* in the image, where is the subject located relative to you?". The second type considers a living being (third-person perspective) within the image as the observer, distinct from both the subject and the object. It includes question templates like, "If you are the *[living being]* in the image, from your perspective, where is the subject located relative to the object?".

Precautions. This part guides annotators in excluding low-quality samples. There are two main precautions to consider: Firstly, the question cannot be correctly answered based solely on the model's prior knowledge without an image. For instance, the question "the book is above the bus" in Figure 1(f) can be answered as "No" without visual input. Secondly, the image must be clear, with the subject or object of the question being easily identifiable.

Table 3: Statistics of our SpatialMQA. The samples in "first-p"(first-person perspective) and "third-p"(third-person perspective) are both from "In-I"(In-image). The latter has fewer samples due to the limited number of images that depict three distinct living entities. "Out-I" means "out-of-image".

	Train	Dev	Test	Total	Ratio	Min.L	Max.L	Avg.L
SpatialMQA	3,780	536	1,076	5,392	100.00	7	34	18.84
Spatial relations								
left of	1,040	148	296	1,484	27.52	7	34	18.39
right of	980	139	279	1,398	25.93	8	32	18.50
in front of	565	80	161	806	14.95	9	33	20.04
behind	529	75	151	755	14.00	7	34	18.61
on/above	353	50	100	503	9.33	8	33	18.30
below	313	44	89	446	8.27	8	32	20.29
Question types								
Out-I	1,513	217	452	2,182	40.00	7	33	15.81
In-I	2,267	319	624	3,210	60.00	12	34	20.91
# first-p	2,136	299	590	3,025	94.24	12	34	20.60
# third-p	131	20	34	185	5.76	18	34	25.91

3.3 Annotation Procedure

To create a high-quality benchmark, we organize a professional team of three annotators, two checkers, and one reviewer. All team members are trained to understand the definition of the spatial coordinate system, six spatial relations, and annotation guidelines. The procedure includes first-round annotation, second-round checking, and third-round review.

First-round annotation. We invite three college students, assigning 10,000 images to each for annotation. According to the guidelines, they write a reasonable question for each image, select options from a predefined set, and mark the correct answer from the options.

Second-round checking. We invite two other college students to simultaneously check the rationality of all samples. Furthermore, each student is assigned an additional task. One student is responsible for checking whether the correct answer to the question can be determined through prior knowledge without images (corresponding to precaution 1). The other student verifies whether the subject or object in the image is clear (corresponding to precaution 2). Samples identified as unqualified by the checkers are returned to annotators with explanations for correction. This process is repeated until a batch achieves 90% accuracy, as determined by the checkers.

Third-round review. A verified batch is given to a main author for double review. The author randomly inspects 20% of the batch samples. Any unqualified annotations are returned to the check team with explanations, allowing them to refine
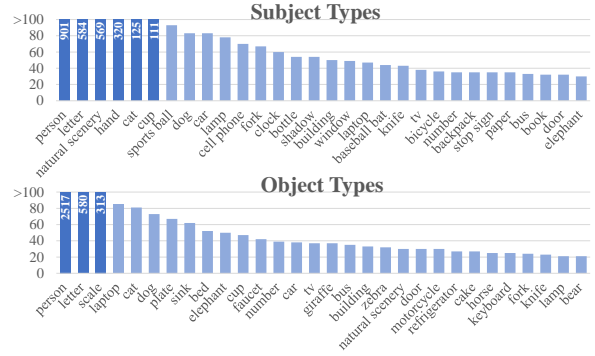

Figure 2: Distributions of subject and object types.

their criteria, which in turn helps standardize the construction team's work. The cycle continues until the batch achieves 95% accuracy. Finally, we obtain 5,392 high-quality samples to form SpatialMQA.

4 SpatialMQA Analysis

Benchmark statistics. As reported in Table 3, SpatialMQA contains 5,392 samples, divided into training, validation, and test sets according to a 7:1:2 ratio. In this benchmark, the questions have a minimum length of 7 words, a maximum length of 34, and an average length of 18.84. Notably, the minimum length of questions in the in-image perspective is 2-3 times longer than the minimum length of questions in the entire benchmark, as these questions typically involve three entities, while other questions generally involve only two.

Diversity of subject and object types in questions. To verify the diversity of questions in our benchmark, we use GPT-4o with in-context learning (ICL) to extract the subjects and objects in questions and classify them into predefined categories. This process is detailed in Appendix B. According to our statistics, there are 113 subject categories and an additional category that includes all subject categories with a sample size of five or fewer, and 84 object categories, along with an additional category that encompasses all object categories with a sample size of five or fewer. Due to the overlap between subject and object types, we have a total of 128 distinct subject and object types. To provide a more intuitive understanding of these types, we present the subject and object types with Top-30 frequency, as shown in Figure 2.

Option combinations. In SpatialMQA, the number of question options varies to ensure they are appropriate for questions. For instance, options like "on/above" and "below" are not suitable

Table 4: Model comparison (%) on our SpatialMQA benchmark. All results are the average of three runs.

Model	Settings	P	R	F1	Acc	Settings	P	R	F1	Acc
Open-source MLLMs										
BLIP-vqa-base	-	32.92	20.86	25.54	26.49	FULL	48.12	31.48	38.06	33.64
BLIP2-opt-2.7B	-	31.31	34.97	33.04	26.86	LoRA	55.20	37.47	44.64	29.93
InstructBLIP-3B	-	37.42	27.47	31.69	28.53	LoRA	44.22	44.80	44.51	42.38
mPLUG-Owl-7B	-	34.30	32.90	33.58	26.49	LoRA	36.05	38.59	37.28	31.88
IDEFICS-9B	-	17.72	25.80	21.00	22.12	LoRA	35.13	36.41	35.76	29.28
LLaVA1.5-7B	-	30.72	31.18	30.95	29.28	LoRA	46.10	44.56	45.32	46.85
SpaceLLaVA	-	35.13	32.58	33.81	31.32	LoRA	47.96	46.18	47.05	**48.14**
Closed-source MLLMs										
Gemini-1.5-flash	0-shot	38.55	35.47	36.95	35.40	2-shot	49.30	35.11	41.01	36.80
	1-shot	51.47	33.46	40.55	36.20	3-shot	51.52	35.82	42.26	38.00
GPT-4o	0-shot	48.62	40.19	44.01	**40.20**	2-shot	48.70	38.36	42.92	38.40
	1-shot	48.04	39.17	43.15	39.00	3-shot	46.76	36.99	41.30	37.80
Other Methods										
Random Choose	-	30.22	27.97	29.05	27.20	-	-	-	-	-
Human	-	98.56	98.40	98.48	**98.40**	Text-only	23.94	24.58	24.26	24.40[3]

for "where is the motorcycle located relative to the car?". Hence, we only include the other four options. Based on the coordinate dimensions, we set the number of options to 2 (two spatial relations in one dimension), 4 (four spatial relations in two dimensions), or 6. According to our statistics, 75% of samples (4036) have 4 options, while 12% (637) and 13% (719) have 2 and 6 options, respectively.

5 Experiments

In this section, we implement state-of-the-art models on our newly constructed SpatialMQA benchmark, aiming at assessing their performance and identifying the underlying challenges.

5.1 Baselines

We mainly select three types of methods: open-source MLLMs, closed-source MLLMs, and others.

Open-source MLLMs. We use BLIP (Li et al., 2022), BLIP2 (Li et al., 2023), InstructBLIP (Dai et al., 2024), mPLUG-Owl (Ye et al., 2023), IDEFICS (Laurençon et al., 2023), LLaVA (Liu et al., 2024), and SpaceLLaVA (Chen et al., 2024) for comparison. Two settings are designed: direct inference and instruction tuning. In the first setting, models directly produce the answer given an image, a question, and multiple options. Note that all models except BLIP receive a task prompt, as described in Appendix C. In the second setting, we use different tuning strategies: full parameter updates for BLIP and parameter-efficient tuning

(LoRA (Hu et al., 2021)) for the other MLLMs. Instruction data is generated by transforming the input and output from the training data, and the task prompt remains consistent with the previous setting.

Closed-source MLLMs. We randomly select 500 samples in SpatialMQA and adopt Gemini-1.5-flash and GPT-4o, two of the most powerful models, for our experiments. For both models, we employ two settings: zero-shot reasoning and few-shot reasoning. In the first setting, we input images, questions, and options, and use a task prompt to guide the MLLM to output answers. Detailed prompts are provided in Appendix D. In the second setting, we use 1-shot, 2-shot, and 3-shot ICL with the same instructions. The ICL examples are randomly selected from the training set and fixed for all samples in the test set.

Other methods. We further design two methods: random selection and manual answering. In the first method, we use a random function to select an answer from the options for each question. In the second method, we randomly select 500 samples from SpatialMQA and invite three college students (different from the annotation team in Section 3.3) to answer the questions. The final answer is determined by majority voting, and if the three students provide different answers, the question is considered incorrect.

Table 5: Results (Acc %) grouped by question types and answer types. Q1, Q2, and Q3 represent the question from the "Out-of-image" perspective, "first-person" , and "third-person" perspectives in images. Ax, Ay, and Az represent answers involving "left of" and "right of" on the X-axis, "in front of" and "behind" on the Y-axis, and "on/above" and "below" on the Z-axis. † and ‡ denote the best few-shot settings in the Main Results, specifically 3-shot and 1-shot, respectively.

Model	Settings	Q1	Q2	Q3	Ax	Ay	Az
Open-source MLLMs							
BLIP-vqa-base	Full	40.93	36.10	52.94	39.65	25.64	28.57
BLIP2-opt-2.7B	LoRA	32.30	28.47	23.53	11.65	49.04	53.97
InstructBLIP-3B	LoRA	44.47	40.68	44.12	36.17	48.72	50.79
mPLUG-Owl-7B	LoRA	37.83	28.14	17.65	17.74	46.47	50.79
IDEFICS-9B	LoRA	33.41	26.95	14.71	15.13	45.51	45.50
LLaVA1.5-7B	LoRA	53.14	40.99	64.71	55.71	29.64	48.13
SpaceLLaVA	LoRA	54.87	42.37	58.82	56.00	51.85	31.41
Closed-source MLLMs							
Gemini-1.5-flash	0-shot	42.73	26.83	50.00	39.17	26.25	41.00
	Few-shot†	48.18	26.83	52.94	49.58	21.88	36.00
GPT-4o	0-shot	44.09	33.74	61.76	37.08	47.50	36.00
	Few-shot‡	45.00	32.52	47.06	38.75	38.75	40.00
Other Methods							
Random Choose	-	30.00	24.80	26.47	25.42	27.50	31.00
Human	Text-only	25.91	23.17	23.53	23.75	25.00	25.00
	-	98.51	98.24	100.00	98.61	97.79	98.68

5.2 Settings and Metrics

The hyperparameter settings for the open-source MLLMs are detailed in Appendix E. These models are executed on a workstation with two NVIDIA A100-PCIE-40GB GPUs. In the experiments, we report four metrics: precision (P), recall (R), F1, and accuracy (Acc).

5.3 Main Results

We perform all baseline methods on our SpatialMQA benchmark. The experimental results are presented in Table 4. From the table, we notice that: 1) All MLLMs perform poorly on SpatialMQA, with significant room for improvement compared to the human accuracy of 98.40%. The best-performing model, SpaceLLaVA with LoRA, achieves only 48.14% accuracy, despite being fine-tuned on LLaVA with a large amount of spatial VQA samples. Notably, LLaVA's visual instruction tuning also involves incorporating coordinate data with bboxes and corresponding captions. This indicates that our SpatialMQA benchmark presents a significant challenge for MLLMs. 2) Among open-source MLLMs, instruction-tuned models excel in spatial relation reasoning compared to those without instruction tuning. For instance, the instruction-tuned SpaceLLaVA shows a 16.82% accuracy improvement over its non-instruction-tuned version.

Among closed-source LLMs, GPT-4o performs best with zero-shot learning, but its accuracy decreases as the number of ICL samples increases. In contrast, Gemini's accuracy improves with more ICL samples. The reasons for these opposing results are explained in "Impact of different ICL examples" of Section 5.4. 3) In other methods, when humans answer questions without images, the accuracy is 24.40% (based on a random selection of 500 samples from the test set), which is comparable to random selection and significantly lower than the accuracy achieved with images. This indicates that our benchmark heavily relies on images to answer questions. In other words, our benchmark rarely includes questions that can be answered solely with prior knowledge. Furthermore, manual annotation reveals that only 3 out of 500 samples could be answered using prior knowledge alone.

5.4 Detailed Analysis

Group analysis of question types and answer types. As mentioned in Section 3.2, question types include "Out-of-image" (denoted as Q1) and "In-image" (further divided into "first-person perspective" (Q2) and "third-person perspective" (Q3)). In addition, we classify the answer types as Ax, Ay, and Az, representing answers involving *left of* and *right of* on the X-axis, *in front of* and *be-*

Table 6: Results (Acc %) for different ICL.

Model	Settings	Alig.	Misalig.
Gemini-1.5-flash	1-shot	36.42	35.16
	2-shot	37.14	35.94
	3-shot	38.28	37.09
GPT-4o	1-shot	39.09	37.75
	2-shot	39.43	36.96
	3-shot	39.88	35.87

Table 7: Impact (Acc %) of images and option counts.

		All	Part
Random	-	17.20	27.20
Gemini-1.5-flash	Q+O	23.20	27.60
	I+Q+O	29.60	35.40
GPT-4o	Q+O	26.40	27.80
	I+Q+O	33.80	40.20

hind on the Y-axis, and *on/above* and *below* on the Z-axis, respectively. The results are listed in Table 5. From the table, we observe that human reasoning abilities in spatial relations are generally consistent across different groups, but all models display significant performance discrepancies within these groups. For instance, human scores for Ax, Ay, and Az are consistently around 98%, while SpaceLLaVA with LoRA exhibits a maximum performance gap of 24.59% in these groups. This suggests that it is essential to improve the model's reasoning abilities in various spatial relations in a balanced manner.

Impact of different ICL examples. We introduce ICL samples for closed-source MLLMs in experiments. To explore the impact of different ICL examples, we divide them into two categories: aligned with the input question type and misaligned. For evaluation, we randomly selected 100 samples for question types Q1, Q2, and Q3 respectively (if a certain category has fewer than 100 samples, we use all available samples). The results are listed in Table 6. From the results, we notice that models with aligned ICL examples outperform those with misaligned ICL examples. For instance, GPT-4o with aligned 3-shot ICL examples improves accuracy by 4.01% over misaligned ones. Notably, the decrease in GPT-4o's spatial relation reasoning ability, mentioned in Section 5.3, may be due to the misalignment of examples with the input question type. In contrast, Gemini's performance improves with more ICL examples in the misaligned setting. This could indicate that Gemini effectively utilizes a wider range of examples to enhance generalization and extract relevant features despite the misalignment.

Impact of images and option counts. We conduct analysis experiments by either removing images (I) in the input or using a fixed count of six options (O). The results are listed in Table 7. From the results, we draw the conclusions: 1) MLLMs with Q+O, when tested with varying options, per-
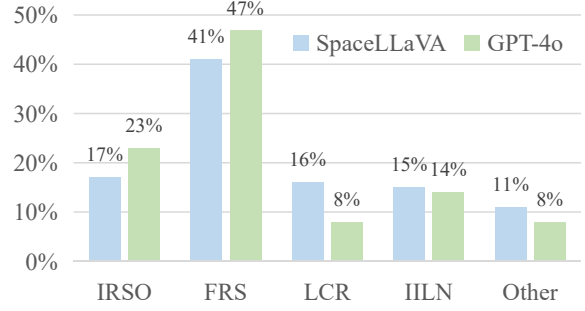

Figure 3: Distribution of error types.

form similarly to random selection and significantly underperform MLLMs with I+Q+O. This indicates that our benchmark heavily relies on image inputs and cannot depend solely on the model's prior knowledge. 2) MLLMs with Q+O still perform significantly better than random selection (17.20%) when given a fixed set of six options. This is because some of the options in this set contradict common sense, allowing the model to exclude them, even without image inputs. This observation is why we remove options that contradict commonsense from our benchmark.

5.5 Error Types

To guide future research in spatial relation reasoning for MLLMs, we analyze 200 error samples produced by SpaceLLaVA and GPT-4o on SpatialMQA. After manual classification, error types are divided into four categories and other errors: (a) incorrect recognition of subjects and objects (IRSO), (b) failure in perspective substitution (FRS), (c) lack of commonsense reasoning ability (LCR), and (d) incorrect identification of spatial relations for letters and numbers (IILN). The error distribution is shown in Figure 3. We observe that FRS errors are the most frequent, with IRSO, LCR, and IILN errors being comparable. To illustrate these error types more intuitively, we provide examples, as shown in Figure 4.
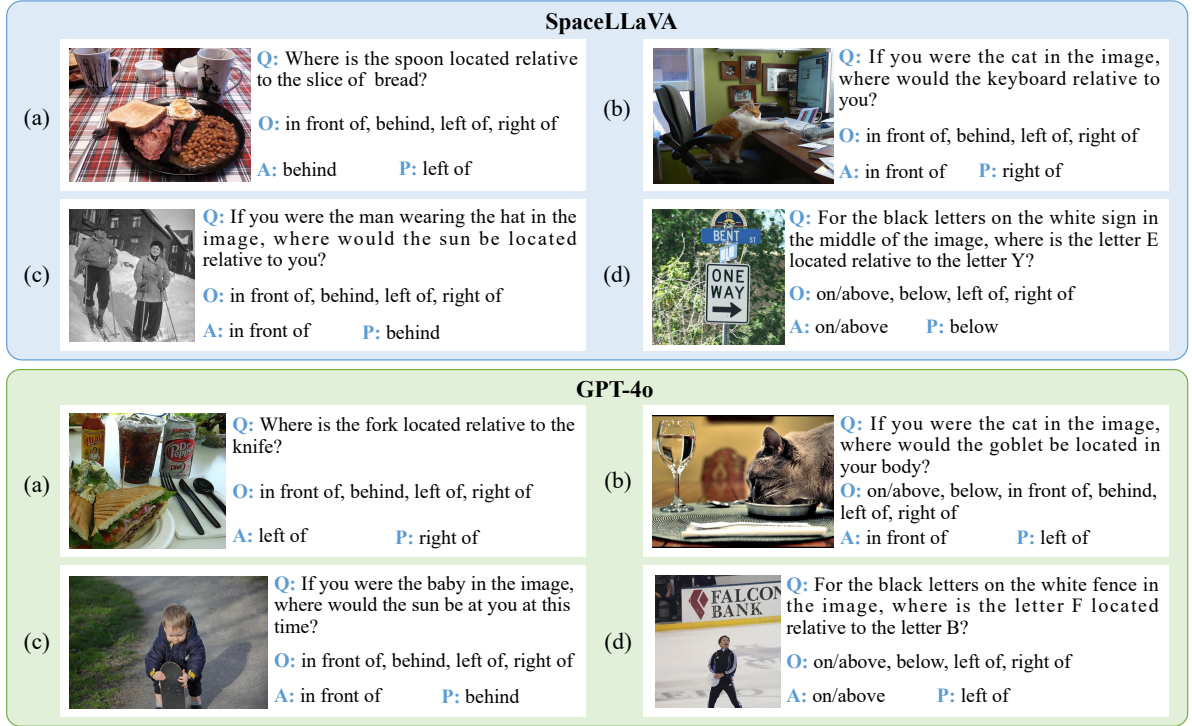
Figure 4: Error examples. (a), (b), (c), and (d) describe examples of the IRSO, FRS, LCR, and IILN types, respectively. "A" and "P" represent the ground truth answer and predicted answer.

6 Related Work

Spatial relation reasoning. Identifying spatial relations between subjects and objects in images is crucial for understanding the world. Benchmarks for this task fall into two main types: those with bboxes and those without. The former are sourced from either synthetic or real-world scenes. CLEVR (Johnson et al., 2017) and Rel3D (Goyal et al., 2020) are typical examples of synthesized benchmarks, but they do not accurately reflect real-world scenes. Hence, several real-scene benchmarks have been proposed, including SpatialVOC2K (Belz et al., 2018), SpatialSense (Yang et al., 2019), SpatialSense+ (Wen et al., 2024), SpatialRGPT-Bench (Cheng et al., 2024), NLVR2 (Suhr et al., 2017), COCO (Lin et al., 2014), and GQA (Hudson and Manning, 2019). However, they still use bboxes, which cause two problems: first, some complex spatial relations can't be fully captured with bboxes (Liu et al., 2023a); second, bboxes make it easier for models to solve tasks without fully understanding the image (Wen et al., 2024). Typical benchmarks for the latter include EmbSpatial-Bench (Du et al., 2024), MME (Fu et al., 2023), SpatialVLM (Chen et al., 2024), EgoThink (Cheng et al., 2023), and VSR (Liu et al., 2023a). The first three focus only

on out-of-image perspectives and do not always annotate samples based on the objective world. Note that SpatialVLM's test set is small and not yet open-sourced. EgoThink is limited to the first-person perspective, with fewer than 100 samples in its spatial reasoning benchmark. While VSR considers different perspectives, only 6% of its data covers them, and some questions can be answered using prior knowledge without images.

Multimodal large language models. With the development of MLLMs, many researchers have applied these models to multimodal downstream tasks. MLLMs can be divided into two categories: closed- and open-source models. Typical closed-source MLLMs include GPT-4o and Gemini. Common methods to adapt these models for multimodal tasks mainly include ICL (Shukor et al., 2023; Liu et al., 2023b) and Chain-of-Thought (CoT) (Zhang et al., 2024; Wang et al., 2024). Typical open-source MLLMs include BLIP2 (Li et al., 2023), LLaVA (Liu et al., 2024), and SpaceLLaVA (Chen et al., 2024). Due to their relatively limited instruction-following capabilities, open-source MLLMs often require instruction tuning for downstream tasks. This tuning can involve full parameter updates or minimal parameter updates, such as LoRA (Hu et al., 2021) and P-tuning v2 (Liu et al., 2021). Despite the promis-

ing progress of current MLLMs, they still perform poorly on our constructed SpatialMQA benchmark.

7 Conclusion

We introduce SpatialMQA, a manually annotated multimodal spatial relation reasoning benchmark based on COCO2017. To address the weaknesses of existing benchmarks, SpatialMQA is constructed without bboxes, involving perspective substitutions based on the objective world and excluding questions that can be answered solely by model's prior knowledge without images. We implement a series of closed- and open-source MLLMs and conducted extensive experimental analyses. The results indicate that SpatialMQA is a challenging benchmark worth further exploration.

Limitations

While SpatialMQA offers a valuable benchmark for evaluating current MLLMs, it has two main limitations: 1) SpatialMQA is created to assess the performance of MLLMs in spatial relation reasoning. To ensure high data quality, we design a manual annotation process, which guarantees a well-constructed and reliable test set. However, this method limits the scale of the training set, making it insufficient for fully fine-tuning MLLMs. Although several automatic annotation tools for spatial relation reasoning are mentioned in (Chen et al., 2024; Cheng et al., 2024; Cai et al., 2024), they are unsuitable for SpatialMQA due to its complex real-world samples from multiple perspectives. 2) SpatialMQA currently covers six basic spatial relations (left of, right of, in front of, behind, on/above, and below), and does not include more complex relations. We focus on these six because experimental results show they already pose significant challenges to current MLLMs. Mastering these fundamental relations is essential before tackling more complex spatial reasoning tasks.

Ethical Statement

Our SpatialMQA benchmark is built upon COCO2017 (Lin et al., 2014), which is licensed under the Creative Commons Attribution 4.0 License. This license allows us to distribute and re-annotate the dataset, as long as the original work is properly cited. Hence, we release SpatialMQA under the CC-BY 4.0 license. Additionally, we have carefully reviewed the benchmark to ensure it contains no harmful content, such as gender bias, racial discrimination, or inappropriate material.

References

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*.

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.

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. 2024. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. 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.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2023. Can vision-language models think from a first-person perspective? *arXiv preprint arXiv:2311.15596*.

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.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394.

Haoxiang Gao, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. 2024. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*.

Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. 2020. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. *Advances in Neural Information Processing Systems*, 33:10514–10525.

Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. 2023. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817.

Zachary Hawes and Daniel Ansari. 2020. What explains the relationship between spatial and mathematical skills? a review of evidence from brain and behavior. *Psychonomic bulletin & review*, 27:465–482.

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*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE CVPR*, pages 2901–2910.

Hugo Laurençon, Daniel van Strien, Stas Bekman, Leo Tronchon, Lucile Saulnier, Thomas Wang, Siddharth Karamcheti, Amanpreet Singh, Giada Pistilli, Yacine Jernite, et al. 2023. Introducing idefics: An open reproduction of state-of-the-art visual language model, 2023. *URL https://huggingface. co/blog/idefics. Accessed*, pages 09–18.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. 2023b. Mmhqa-icl: Multimodal in-context learning for hybrid question answering over text, tables and images. *arXiv preprint arXiv:2309.04790*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Michael J Proulx, Orlin S Todorov, Amanda Taylor Aiken, and Alexandra A de Sousa. 2016. Where am i? who am i? the relation between spatial cognition, social cognition and individual differences in the built environment. *Frontiers in psychology*, 7:158846.

Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2023. Beyond task performance: evaluating and reducing the flaws of large multimodal models with in-context-learning. In *The Twelfth International Conference on Learning Representations*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223.

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

Junjie Wang, Wei Li, Yinjian Wang, Ran Tao, and Qian Du. 2023. Representation-enhanced status replay network for multisource remote-sensing image classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170.

Chuan Wen, Dinesh Jayaraman, and Yang Gao. 2024. Can transformers capture spatial relations between objects? *arXiv preprint arXiv:2403.00729*.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. 2019. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2051–2060.

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

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*.

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

A Question Templates

In Section 3.2, we design three types of questions. For each type, we manually define several question templates, as listed in Table 8. Q1, Q2, and Q3 indicate that the sample's question type is "Out-of-image", the first-, and third-person perspective of "In-image", respectively.

Table 8: Question Template.

	Question Template
Q1	Is xx located to the left or right of xx? Which side of xx is xx located on? Where is xx located relative to xx?
Q2	If you are xx in the image, is xx located to your left or right? If you are xx in the image, which side of xx is xx located on? If you are xx in the image, where is xx located relative to you?
Q3	If you are xx in the image, from your perspective, is xx located to the left or right of xx? If you are xx in the image, from your perspective, which side of xx is xx located on? If you are xx in the image, from your perspective, where is xx located relative to xx?

B Statistics of Subject and Object Types

In Section 4, we use GPT-4o with ICL to extract the subjects and objects in questions and classify them into predefined categories. The process is as follows. First, we adopt GPT-4o with ICL to extract the subject and object of each question in SpatialMQA. Second, we randomly select 500 samples from the entire benchmark and manually define common types, in addition to the original 80 types from COCO2017, resulting in a total of 90 types. Third, we employ GPT-4o with ICL to classify every subject and object into these 90 types. Finally, samples that are not classified into predefined types are manually categorized into new types.

C Details of Open-source MLLMs

In Section 5.1, we consider open-source MLLMs as baseline models. The task prompts and instruction data format of these models are presented in Tables 9 and 10.

D Details of Closed-source MLLMs

In Section 5.1, we consider closed-source MLLMs as baseline models. The task prompts of these models are listed in Table 11.

Table 9: Task prompts for open-source MLLMs.

Models	Task prompt
BLIP	Input: Image: <image>, Question: {question}, Options: {options}. \n Output:
BLIP2, InstructBLIP, IDEFICS, mPLUG-Owl, LLaVA, SpaceLLaVA	You are currently a senior expert in spatial relation reasoning. \n Given an Image, a Question, and Options, your task is to answer the correct spatial relation. Note that you only need to choose one option from all options without explaining any reason. \n Input: Image: <image>, Question: {question}, Options: {options}. \n Output:

Table 10: Instruction data format for open-source MLLMs.

Models	Instruction
BLIP	Input: Image: <image>, Question: {question} Options: {options}. \n Output: {answer}
BLIP2, InstructBLIP, IDEFICS, LLaVA, SpaceLLaVA	You are currently a senior expert in spatial relation reasoning. \n Given an Image, a Question, and Options, your task is to answer the correct spatial relation. Note that you only need to choose one option from all options without explaining any reason. \n Input: Image: <image>, Question: {question}, Options: {options}. \n Output: {answer}
mPLUG-Owl	The following is a conversation between a curious human and an AI assistant. \n Human: <image> \n Human: You are currently a senior expert in spatial relation reasoning. \n Given an Image, a Question, and Options, your task is to answer the correct spatial relation. Note that you only need to choose one option from all options without explaining any reason. \n Input: Image: <image>, Question: {question}, Options: {options}. \n Output: \n A: {answer}

E Hyperparameter Settings

Details of the hyperparameter settings for open-source MLLMs are presented in Table 12.

F Annotation Tool

To enhance annotation efficiency, we develop a tool used for annotating (Figure 5) and checking (Figure 6) samples in SpatialMQA, as well as answering (Figure 7) questions in the test set for evaluators. Each volunteer was compensated at a rate of $17 per hour.

Table 11: Task prompts for closed-source MLLMs.

|	**Task prompt**
Zero-shot	You are currently a senior expert in spatial relation reasoning. \n Given an Image, a Question, and Options, your task is to answer the correct spatial relation. Note that you only need to choose one option from all options without explaining any reason. \n Input: Image: <image>, Question: {question}, Options: {options}. \n Output:
Few-shot	You are currently a senior expert in spatial relation reasoning. \n Given an Image, a Question, and Options, your task is to answer the correct spatial relation. 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 spatial relation reasoning task: \n Example1: Input: Image: <image> \n Question: For the clock in the image, does the hour hand point above or below the 9 scales?, Options: on/above; below. \n Output: above. \n Example2: ... \n Example3: ... \n Input: Image: <image> \n Question: {question}, Options: {options}. \n Output:
Text-only	You are currently a senior expert in spatial relation reasoning. \n Given an Image, a Question, and Options, your task is to answer the correct spatial relation. Note that you only need to choose one option from all options without explaining any reason. \n Input: Question: {question}, Options: {options}. \n Output:
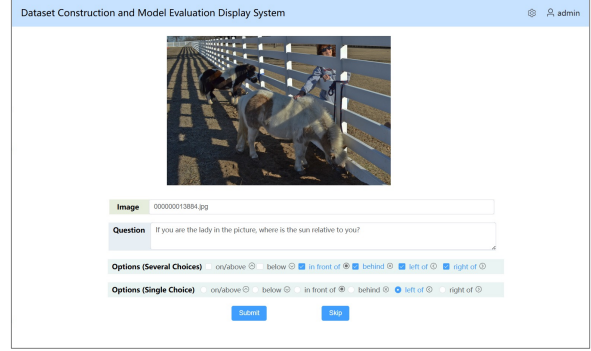
Figure 5: First-round annotation page in our tool.
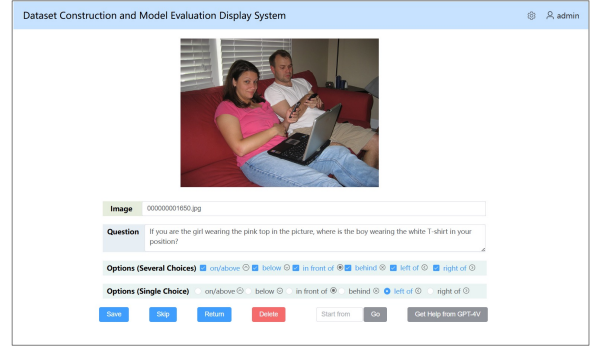
Figure 6: Second-round checking and third-round review pages in our tool.

Table 12: Hyperparameter settings for open-source MLLMs. "Ep", "BS", "ES", "LR", "Opt", "LR. S", "PAW8", "ExpLR" and "LD" stand for "Epochs", "Batch Size", "Early Stop", "Learning Rate", "Optimizer", "LR Schedule", "Paged_Adamw_8bit", "ExponentialLR", and "Linear Decay", respectively.

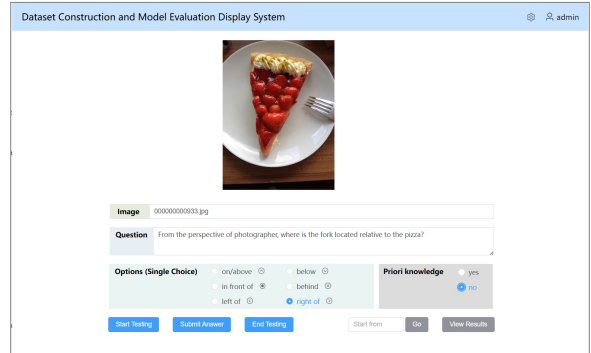| Model	Ep	BS	ES	LR	Opt	LR. S
BLIP	30	8	5	6e-7	AdamW	ExpLR
BLIP2	30	8	5	4e-5	AdamW	ExpLR
InstructBLIP	30	8	5	4e-5	AdamW	ExpLR
mPLUG-Owl	10	8	-	5e-5	AdamW	LD
IDEFICS	10	8	-	2e-4	PAW8	LD
LLaVA	10	8	-	2e-4	AdamW	Cosine
SpaceLLaVA	10	8	-	2e-4	AdamW	Cosine

Figure 7: Human evaluation page in our tool.