An Empirical Analysis on Spatial Reasoning Capabilities of Large Multimodal Models

Fatemeh Shiri[♡] **Xiao-Yu Guo**[*♢] **Mona Golestan Far**[♡]
Xin Yu[♣] **Gholamreza Haffari**[♡] **Yuan-Fang Li**[♡]

[♡]Department of Data Science & AI, Monash University

[♢] Australian Institute for Machine Learning, University of Adelaide

[♣]School of Electrical Engineering and Computer Science, University of Queensland

`{first.last}@monash.edu`

`xiaoyu.guo@adelaide.edu.au`

`xin.yu@uq.edu.au`

Abstract

Large Multimodal Models (LMMs) have achieved strong performance across a range of vision and language tasks. However, their *spatial* reasoning capabilities are under-investigated. In this paper, we construct a novel VQA dataset, Spatial-MM, to comprehensively study LMMs' spatial understanding and reasoning capabilities. Our analyses on object-relationship and multi-hop reasoning reveal several important findings. Firstly, bounding boxes and scene graphs, even synthetic ones, can significantly enhance LMMs' spatial reasoning. Secondly, LMMs struggle more with questions posed from the human perspective than the camera perspective about the image. Thirdly, chain of thought (CoT) prompting does not improve model performance on complex multi-hop questions involving spatial relations. Lastly, our perturbation analysis on GQA-spatial reveals that LMMs are much stronger at basic object detection than complex spatial reasoning. We believe our new benchmark dataset and in-depth analyses can spark further research on LMMs spatial reasoning. [1]

1 Introduction

Large Multimodal Models (LMMs) have shown impressive generalization ability on several vision and language tasks. Several recent works, however, showed that these models lack spatial understanding (Tong et al., 2024; Li et al., 2023a; Liu et al., 2023e; Lei et al., 2024a; Prasad et al., 2023). As can be seen in Figure 1, multimodal LLMs, including GPT-4o, often fail to answer questions from a human perspective within an image. The focus of this work is to study the understanding of *spatial* relations by top-performing LMMs. Moreover, we go beyond evaluating only the final answers to directly analyzing the intermediate reasoning steps

Figure 1: benchmarking the spatial reasoning capabilities of GPT-4o (Achiam et al., 2023) (Date accessed: June 12, 2024). Text in red and green signifies an incorrect and ground-truth answers, respectively. The accuracy of GPT-4o in answering questions related to the human's viewpoint in the image is only 27.5%.

generated by chain of thought (CoT) prompting in multi-hop visual question-answering (VQA) tasks. We ground LMMs ' reasoning steps into a scene graph format and verify whether they form a valid path.

More specifically, we ask the following questions: (i) What spatial relations are missed by models, and why it happen? (ii) How can additional symbolic visual information, such as bounding boxes or scene graphs, improve the performance of LMMs? Which of these symbolic information is more useful, and how can they be integrated in the reasoning process effectively? (iii) How does the questions complexity affect LMMs in handling spatial relations? (iv) How does the reasoning path of LMMs behave when they fail to answer a multi-hop question? Is the failure due to incorrect spatial reasoning or non-spatial reasoning?

To address these questions, we construct Spatial-MM, a novel, challenging dataset, and comprehensively LMMs spatial reasoning capabilities from different angles. We analyze four top-performing LMMs on Spatial-MM and GQA-spatial (Kamath et al., 2023) benchmarks to identify problems with visual question answering (VQA) evaluation methodology. Our comprehensive analyses reveal a number of important insights

---

[1]Spatial-MM benchmark is available at: `https://github.com/FatemehShiri/Spatial-MM`

that point to future research directions. Our contributions can be summarized as follows.

• We present a new, challenging spatial-aware benchmark that incorporates a variety of spatial relationship types, accounting for both human and camera perspectives.

• Our coprehensive empirical analyses show that: (i) bounding boxes and scene graphs, even synthetic ones, can significantly enhance LMMs' spatial reasoning, (ii) LMMs struggle more with questions posed from the human perspective than the camera perspective about the image, (iii) chain of thought (CoT) prompting does not improve model performance on complex multi-hop questions involving spatial relations, and (iv) LMMs are much stronger at basic object detection than complex spatial reasoning.

2 Related Work

2.1 Large Multimodal Models

Pre-trained on large scale of text corpus, Large Language Models (LLMs) (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020; Zheng et al., 2023b) can easily reach more than billions of parameters and show great capacity on natural language comprehension, text completion and generation, zero-shot transfer and in-context learning. However, traditional LLMs usually take text as input and output, and lack the ability of understanding other modalities, like image, video or audio.

To tackle this issue, connecting multimodal encoder/decoder with LLMs, Large Multimodal Models (LMMs) can integrate multiple data modalities and overcome the limitations of text-only LLMs. LMMs are utilized to address various different tasks: image-text understanding (Liu et al., 2023c,a; Li et al., 2023c), video-text understanding (Lin et al., 2023; Li et al., 2023d; Maaz et al., 2023), and even multimodal generation (Peng et al., 2023; Lv et al., 2023; Zheng et al., 2023a).

In this paper, image-text understanding is our main focus, because it is already quite hard for LMMs to understand the spatial relation in normal images, not to mention videos.

2.2 Spatial relationship Benchmarks

Though LMMs have demonstrate remarkable performance in various benchmarks, they still have weaknesses on understanding spatial relationships, such as distinguish "left" from "right" between two objects presented in one image.

Kamath et al. (2023) curate a new What'sUP benchmark to quantify model performance on understanding spatial relationships. The evaluation results show not only the pretraining corpus contains little spatial-related data to learn from, but also LMMs perform limited on this spatial benchmark. Chen et al. (2024) identify the similar problem in 3D spatial relations. By generating a large scale of spatial annotations, they increase the LMMs spatial reasoning skill by a large margin. Instead of increasing the scale of pre-training data, Lei et al. (2024b) deal with spatial relationships from the model perspective. They propose SCAFFOLD prompting that overlays a dot matrix onto the image as visual information anchors, which demonstrates better performance in spatial reasoning over GPT-4 Vision with the textual CoT prompting.

In this paper, we propose a new spatial reasoning benchmark Spatial-MM, covering a diverse spatial relationships more than What'sUP and containing less noisy or ambiguous annotations.

2.3 Multi-hop Reasoning

Chain of thought prompting (Wei et al., 2022a) demonstrates remarkable multi-hop reasoning capability of LLMs by eliciting step-by-step reasoning paths. Least-to-most prompting (Zhou et al., 2022) further shows the feasibility of conducting decomposition and multi-hop reasoning, which happens on the decoder side together with the answer prediction procedure. Furthermore, Kil et al. (2024) improves multimodal multi-hop reasoning in VQA. Using answer prediction-guided CoT, the II-MMR model finds a reasoning path to reach the answer.

Our new benchmark dataset Spatial-MM contains spatial multi-hop reasoning questions, covering at least two reasoning steps for each ground-truth reasoning path. To the best of our knowledge, it is the first multimodal multi-hop benchmark that pays attention to the evaluation on the reasoning path.

3 The Spatial-MM Benchmark

We seek to study the spatial reasoning capability gap between humans and LMMs. Based on the observation that existing benchmarks only partially investigate the spatial reasoning capabilities of LMMs (Li et al., 2023b; Liu et al., 2023f), we introduce a novel benchmark, Spatial-MM, which includes two subsets: Spatial-Obj and Spatial-CoT. Spatial-Obj features multiple-choice questions that

focus on the spatial relationships between one or two objects in an image, while Spatial-CoT offers open-ended multi-hop questions.

3.1 Spatial-Obj

Spatial-Obj is a carefully curated benchmark that contains 2,000 multiple-choice questions with the aim of assessing LMMs' spatial reasoning of one or two objects in a given image. With natural images downloaded from the Internet, we carefully selected images to construct diverse challenging multiple-choice questions, including both yes/no questions and wh-type questions.

The dataset is constructed with two rounds of annotations. In the first round of annotation, we divided the images among three annotators and tasked them with selecting one or two objects in each image and compose a question-answer (QA) pair including a spatial relationship. Annotators were provided with question templates with objects placeholders, which they could use or customize according to their preference. In the second round of annotation, we released batches of 200 QA pairs with their corresponding images. Another 10 annotators were tasked with reviewing these QA pairs to verify whether they were correct or incorrect/ambiguous. Corrections were made based on their feedback.

This dataset covers 36 of the most commonly used spatial relationships (Marchi Fagundes et al., 2021), including *right, left, attached to, touching, back, bottom, ahead, forward, backward, down, facing towards, facing away, top, beneath, beside, side, behind, under, on, in, front, below, above, over, middle, between, inside, outside, bottom right, bottom left, top right, top left, corner, close to, next to, near*. Moreover, we used GPT-4o to categorise Spatial-Obj into visual patterns such as "object localization", "orientation and direction", "viewpoints" and "positional and relational context", which pose significant challenges for LMMs. The prompt is listed in Appendix 7. Examples of these patterns can be seen in Figure 2.

3.2 Spatial-CoT

Chain of Thought (CoT)-style prompting has been demonstrated to significantly improve LLMs' reasoning capabilities. However, recent investigations on knowledge graph question answering and mathematical reasoning show that discrepancies exist between the answer and the corresponding reasoning paths through CoT prompting (Nguyen et al.,

2024; Zhou et al., 2024). That is, while, by using CoT to producing reasoning paths, the LLM can produce the correct answer, the generated reasoning paths are not always correct. This disparity is a form of hallucination, and renders the LLMs' reasoning less trustworthy.

To enable the study of faithfulness in LMMs' spatial reasoning in the CoT style, we curate Spatial-CoT, a dataset of multi-hop question-answer pairs. The QA pairs are generated by prompting GPT-4o with a given image and a set of in-context examples. Images are sourced from Internet. In total, 800 multi-hop QA pairs were generated, and we filtered out 178 that did not include at least one of the 36 spatial relationships listed in Section 3.1 above. We then employed human annotators to select reasonable and meaningful multi-hop QA pairs that require at least two reasoning steps to reach the final answer. An additional 312 QA pairs were manually discarded for lacking the complexity needed for multi-hop QA pairs. Ultimately, Spatial-CoT includes 310 spatial-aware multi-hop QA pairs. The prompt, including instructions for generating multi-hop QA, is provided in Section A of the appendix.

Spatial-aware reasoning paths. In knowledge graph question answering, triplets are considered reasoning steps (hops) (Nguyen et al., 2024). Leveraging this idea, in VQA tasks, we utilize scene graphs, including object relations or attribute, as reasoning steps that lead to the answer. It is important to note that current VQA benchmarks (e.g. GQA (Hudson and Manning, 2019)) with ground-truth scene graphs lack diverse perspective information and only incorporate spatial relationships from the camera's viewpoint. For Spatial-CoT, as no ground-truth paths (or scene graph) exist, to evaluate the spatial reasoning abilities of LMMs, we generate the reasoning paths for each question. We carefully instruct GPT-4o with a set of in-context examples to generate the initial draft of the spatial-aware reasoning path. Subsequently, three annotators were tasked with evaluating (i.e. keeping, removing, adding, or modifying) the generated steps for each question. Additionally, the annotators were asked to tag the steps as S for spatial hops such as "person in front of car" and NS for non-spatial hops such as "woman holding a fork", respectively. In total, 67% of reasoning steps are tagged as S and the rest 33% are tagged as NS, suggesting that many questions require reasoning of spatial relationships between key objects to be

Orientation and Direction

Is the giraffe with the shorter neck is facing his left?

A: yes
B: no

🟢 A. yes

Viewpoint

From which viewpoint can the mug held by the person be seen?

A: side   B: top
C: bottom D: front

🟢 D. front

Positional & Relational Context

Where are the cows in from the man's perspective?

A. behind
B. left
C. front
D. right

🟢 D. right

Object Localization

Where is the dog from the camera's perspective?

A. left.
B. middle
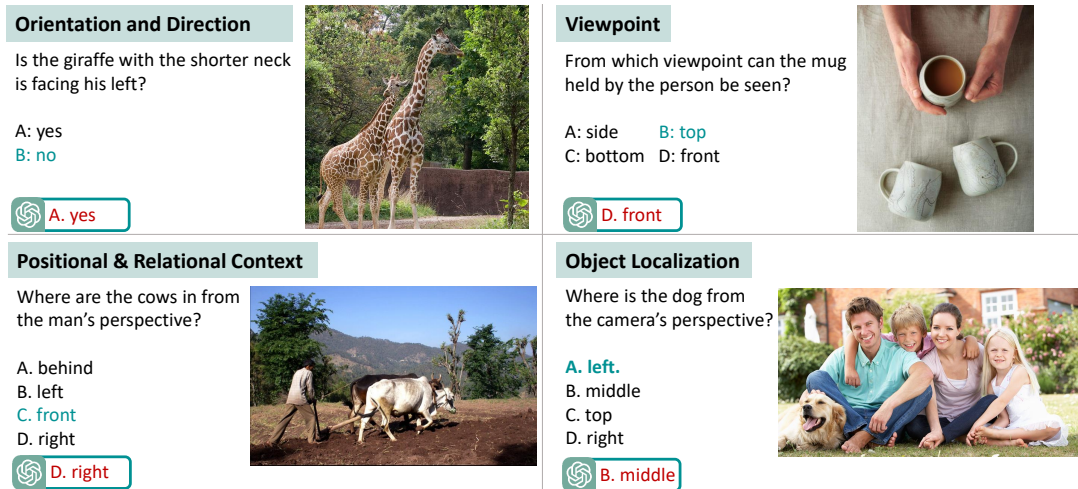C. top
D. right

🟢 B. middle

Figure 2: VQA examples from our Spatial-MM that encompass a range of challenging visual patterns..

answered.

Each reasoning path includes two or more hops, with at least one being a spatial hop. Figure 3 shows some examples of multi-hop QAs with their corresponding Reasoning path. As shown in Table 3, among the questions, 34% have a reasoning path with two hops, 39% have three hops, and 27% require four or more hops to answer.

3.3 Human performance

We evaluated human performance using our benchmark to maintain the quality of annotations. We sampled 300 data points from Spatial-Obj and 100 samples from Spatial-CoT. Annotators were consulted to assess whether the correct option is clearly identifiable or if there's any potential for confusion. This human performance assessment yielded scores of 98% and 99% on Spatial-Obj and Spatial-CoT datasets, respectively.

4 Data Enrichment to Improve Spatial Reasoning

Intuitively, providing LMMs with additional spatial-aware visual grounding data would help improve their spatial reasoning performance. To test this hypothesis, we describe two pipelines for generating different types of visual grounding data. The first pipeline consists of two stages: (I) key object extraction and (II) bounding box generation. The second pipeline consists of three stages: (I) key object extraction, (II) spatial-aware caption generation related to the key objects, and (III) spatial-aware scene graph generation. Please refer to Figure 6 in Appendix C for detailed illustrations of our pipelines. Furthermore, we describe Reasoning

Question: Where is the pink bear from the perspective of the woman wearing the floral dress?
Answer: right
Reasoning: [

    {bear1 is pink},

    {bear1 is on the left},

    {woman1 wearing floral dress},

    {women1's right is image's left},

    {bear1 on the right from women1's perspective}]

Figure 3: An example of generated reasoning steps for a multi-hop question.

path generation for the multi-hop questions.

Bounding box generation. The input of the pipeline is an image I paired with a multiple-choice question Q. We first extract key objects, $[object_1, object_2, ..]$ from the given question. Next, we prompt GPT-4o to provide the bounding boxes of the key objects in the image. Each bounding box is specifically represented as a tuple $[x_{min}, y_{min}, x_{max}, y_{max}]$, where x_{min} and y_{min} are coordinates of the top-left corner and x_{max} and y_{max} are coordinates of the bottom-right corner. We present the prompt for bounding box generation in Appendix 7, and examples of synthesized bounding boxes for the key objects along with the

ground-truth bounding boxes in Appendix D.

Scene graph generation. Similar to the first stage of bounding box generation, we extract key objects from the given question. We then prompt GPT-4o to generate a spatial-aware caption by considering the location, direction, orientation, and spatial relations of the key objects in the image. Finally, given the question, spatial-aware caption, key objects, and image, GPT-4o is prompted to extract spatial relation, orientation, and direction of the key objects in the image. The prompt for scene graph generation is listed in the Appendix 7.

Reasoning path generation. Previous research on evaluating LMMs has solely focused on final answer accuracy, neglecting the correctness of the generated reasoning paths (Lei et al., 2024a; Kil et al., 2024). In this work, we delve deeper into the spatial reasoning path of LMMs in multi-hop VQA by utilizing the ground-truth reasoning steps we generated for multi-hop questions in Section 3.2. To address the challenge of evaluating unstructured reasoning steps, we carefully design a prompting strategy to instruct LMMs to output the reasoning steps in a scene graph format. This enables us to validate against ground-truth reasoning steps. Figure 3 shows an example of generated reasoning paths for an image.

5 Experiments

5.1 Dataset

We use the following datasets in our experiments:
• **Spatial-MM** contains **Spatial-Obj** described in §3.1 and **Spatial-CoT** described in §3.2.
• **GQA-spatial** (Kamath et al., 2023) is sourced from GQA (Hudson and Manning, 2019), where each image is paired with two caption options, which contain opposite spatial relationships, i.e. *prepositions*. Depending on the number of *object(s)* identified in the image, GQA-spatial contains 1-object and 2-object caption options. One instance from GQA-spatial is shown in the "standard" row of Figure 4: one image and two captions make use of opposite *preposition*s: *front* and *behind*.[2]

5.2 LMMs

We evaluate the spatial reasoning capability of the following four LMMs: **LLaVA-1.5-7B** (Liu et al., 2023b), **GPT-4 Vision** (OpenAI, 2023), **GPT-4o** (OpenAI, 2024), **Gemini 1.5 Pro** (Gemini Team,

---

[2]There are three pairs of opposite spatial relationships in GQA-spatial: {*left, right*}, {*front, behind*} and {*top, bottom*}.

2024), and **MiniGPT-v2 (7B)** (Chen et al., 2023). Please refer to Appendix A for details.

5.3 Evaluation Metrics

To evaluate model performance on final answer prediction, we adopt the standard Accuracy metric Hudson and Manning (2019). For evaluating Spatial-CoT, the exact-match metric is insufficient for measuring performance fully due to the open-ended nature of the answers. LMMs frequently produce paraphrases or alternative expressions that convey the same underlying meaning. Therefore, we initially verify if the predicted answer exactly matches the gold answer. If it does not, following (Lyu et al., 2024; Wang et al., 2023), we use GPT-4o as a judge to determine whether the predicted answer holds the same semantic meaning as the gold answer.

5.4 The Effect of Bounding Boxes and Scene Graphs

For both Spatial-Obj and GQA-spatial benchmarks, the input is an image paired with a multiple-choice question. Table 1 presents the results on Spatial-Obj and GQA-spatial, comparing LMMs' performance on images only and with additional information (i.e. bounding boxes and scene graphs). A number of important observations can be made from the table. (I) It shows that the additional bounding box information serves as visual anchors that enhance LMMs' spatial reasoning performance for LMMs. When given the synthesize bounding boxes information on GQA dataset, the LMMs improve their answer accuracy significantly, with the average increase over "stan" of 33 points. However, spatial related questions require a profound semantic understanding of object relationships within the scene. By incorporating spatial-aware scene graphs of key objects, the performance of LMMs improves significantly on both datasets. Specifically, by including the synthesized SG in the prompt, GPT-4 vision's accuracy increases by 9.28% and 8.92% for one-object and two-object questions on the Spatial-Obj dataset. (II) Surprisingly, the LMMs perform better with synthesized bounding boxes than with ground-truth bounding boxes. This may be attributed to the high quality and diversity of synthesized bounding boxes generated by GPT-4 vision. (III) On Spatial-Obj, scene graphs further enhance model performance, while on GQA-spatial, bounding box information seems to be more useful. We leave a deeper analysis of these two observations

Prompt	GPT-4 vision			Gemini			LLaVA-1.5			MiniGPT-v2		
	1_obj	2_obj	all	1_obj	2_obj	all	1_obj	2_obj	all	1_obj	2_obj	all
Spatial-Obj												
stan	50.80	52.85	52.18	46.63	41.53	43.18	51.59	43.58	46.18	45.73	35.83	39.99
stan+bbox$_{synth}$	**71.34**	60.21	**63.82**	50.81	45.55	47.26	**56.97**	44.62	48.63	**53.61**	40.87	**44.93**
stan+SG$_{synth}$	60.08	**61.77**	61.22	**53.42**	**55.45**	**54.79**	51.60	**50.79**	**51.05**	49.44	**41.39**	43.95
GQA-spatial												
stan	66.46	15.81	56.30	19.40	16.35	18.79	27.16	15.12	24.75	22.14	8.63	19.43
stan+bbox$_{gt}$	79.69	40.89	71.91	33.74	37.84	33.76	68.19	42.96	63.13	33.45	21.69	31.09
stan+bbox$_{synth}$	**85.86**	64.95	**81.67**	**42.11**	48.00	43.29	**80.86**	50.52	**74.78**	50.26	39.99	48.20
stan+SG$_{gt}$	71.60	**70.69**	71.42	41.36	**55.21**	**44.14**	55.40	36.90	51.69	39.67	31.69	38.07
stan+SG$_{synth}$	64.90	55.40	62.99	24.77	49.25	29.68	36.99	46.55	38.91	51.77	33.9	48.19

Table 1: The accuracy of models in the Standard (stan) setting, augmented with synthesized bounding boxes (bbox$_{synth}$), scene graphs (SG$_{synth}$), ground truth bounding boxes (bbox$_{gt}$) and ground truth scene graphs(SG$_{synth}$). The synthetic bounding boxes and scene graphs are produced by GPT-4o. **Bold**: Best results within each dataset.

	Human's pers.	Camera's pers.	Overall
Spatial-Obj			
Distribution	31%	69%	100%
GPT-4o	43.75	73.6	64.89
GPT-4 vision	42.56	54.81	51.23
LLaVA	37.70	46.04	43.61
Gmini	32.92	44.07	40.82
MiniGPT-v2	31.67	40.75	37.94
Spatial-CoT			
Distribution	46%	54%	100%
GPT-4o	27.50	63.25	46.81
GPT-4 vision	33.05	50.14	42.28
LLaVA	22.91	40.47	32.39
Gmini	25.81	46.29	36.87
MiniGPT-v2	22.08	39.77	31.63

Table 2: Results of human and camera perspectives.

Num of hops		2-hop	3-hop	$\geq$ 4-hop	All
Hop Distribution		34%	39%	27%	100%
GPT-4o	stan	76.47	60.00	42.86	60.78
	stan+COT	76.47	55.54	46.89	56.91
	stan+bbox$_{synth}$	**82.35**	55.0	35.71	58.82
	stan+SG$_{synth}$	73.33	**75.0**	35.71	63.27
GPT-4v	stan	58.82	55.0	50.0	54.9
	stan+COT	50.0	60.0	**60.0**	56.86
	stan+bbox$_{synth}$	56.25	50.0	50.0	52.0
	stan+SG$_{synth}$	81.25	**75.0**	35.71	**66.0**
LLaVA	stan	47.06	50.0	50.0	49.02
	stan+COT	35.29	55.0	21.43	39.22
	stan+bbox$_{synth}$	47.09	50.0	35.71	46.14
	stan+SG$_{synth}$	52.94	60.0	50.0	54.9
Gmini	stan	52.94	50.0	42.86	49.02
	stan+COT	33.33	29.41	36.84	33.33
	stan+bbox$_{synth}$	56.25	30.0	28.57	38.0
	stan+SG$_{synth}$	68.75	55.0	57.14	60.0

Table 3: We evaluated the accuracy of LMMs on different numbers of hops on our challenging multi-hop benchmark, Spatial-CoT.

to future work.

> **Finding 1:** Bounding boxes and scene graphs enhance LMMs 's ability to excel in visual reasoning tasks. Bounding boxes are more effective in one-object questions, while scene graphs are more helpful in two-object ones.

5.5 The Effect of Human and Camera Perspectives

Questions from previous benchmarks (Kamath et al., 2023) are often evaluated from the camera's perspective, i.e., **outside** the image. How would LMMs behave if the question is asked from other angles, such as from the human perspective in the image, i.e., **inside** the image, is under-explored. This problem motivates us to generate different questions with distinct prompts "from camera's perspective", "from human's perspective" respectively, from original questions, and further understand the spatial reasoning capabilities of LMMs.

In Table 2, we show different model performance from either human's or camera's perspective on Spatial-Obj and Spatial-CoT. Though all these LMMs receive explicit prompt "from human/camera's perspective", they all show signif-

icant performance drop on human's perspective, when compared to the camera's perspective. Especially, while GPT-4o outperforms all other models from camera's perspective, it is still not able to understand from human's perspective and show the maximum drop on Spatial-CoT of more than 30 points.

> **Finding 2:** LMMs excel at understanding the scene from the camera's perspective. However, their performance declines significantly when questions are posed from the human perspective within the image.

5.6 Analysis based on Complex Multi-hop VQAs

The effect of the number of hops to the answer. Table 3 compares chain of thought prompting with other approaches on the Spatial-CoT dataset. As can be seen, the conventional CoT prompting (Wei et al., 2022b) may not be as effective for complex VQA tasks as it is for NLP tasks. Indeed, we can

Reasoning type		Spatial			Non-spatial		
Distibution of steps		67%			33%		
Models	Prompting	P	R	F1	P	R	F1
GPT-4o	stan	79.28	63.77	70.28	90.48	95.02	92.68
	stan+SG$_{synth}$	80.81	68.38	74.07	91.43	96.00	93.66
Gemini	stan	66.99	69.70	68.32	86.27	89.80	88.01
	stan+SG$_{synth}$	80.21	64.38	71.43	90.10	92.86	91.46
LLaVA	stan	66.02	71.58	68.96	85.62	92.71	89.02
	stan+SG$_{synth}$	77.23	67.83	72.22	82.41	93.68	87.68

Table 4: Evaluation of Spatial and Non-spatial reasoning steps on our Spatial-CoT benchmark.

Reasoning Path	Correct	Incorrect	
Reasoning Type	None	Spatial	Non-spatial
GPT-4o	0.53	92.55	6.91
GPT-4 vision	1.18	91.76	8.24
LLaVA	1.99	89.40	12.58
Gmini	1.32	90.73	11.26
Average (%)	1.25	91.11	9.75

Table 5: Evaluation of reasoning path on multi-hop questions which LMMs fail to answer correctly.

observe that even the strong LMMs such as GPT-4o, Gemini-Pro and LLaVA-1.5 perform worse when employing CoT reasoning, compared to standard prompting. Therefore, the rationales produced by the conventional CoT may not align well with the reasoning path needed to arrive at the answer.

On the contrary, including visual grounded information such as bounding boxes information is effective when the number of reasoning hops are ≤ 3. In particular, Table 3 highlights the effectiveness of scene graphs in answering multi-hop spatial questions across all the models. Multi-hop questions demand a deep semantic understanding of object attributes and relationships within the scene. Therefore, scene graphs are essential for enhancing LLMs' semantic visual understanding. Moreover, across different models and various experiment settings in Table 3, it can be seen that with the increase in the number of the hops, accuracy of the final answer drops. For instance, the average accuracy of questions with ≥ 4 hops is 12% and 7.5% less than that for questions with 2 and 3 hops, respectively.

Correctness of the reasoning path. We ground LMMs' reasoning steps into a scene graph format and verify whether they form a valid path. Given the question and predicted answer, we carefully design prompts to instruct LMMs to output the reasoning steps in a scene graph format, i.e., including object relations and/or attributes. This enables us to validate the LMMs' reasoning steps against ground-truth reasoning paths (Section 4). To evaluate the reasoning path in multi-hop questions, computing
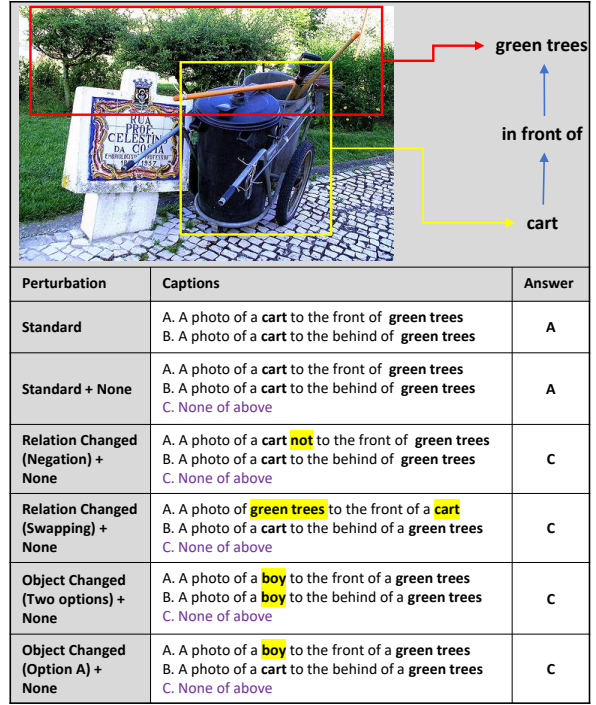

Figure 4: Examples of caption perturbation.

the *semantic match* for all the generated steps in the path. We further analyze the types of reasoning steps required in multi-hop questions.

Table 5 presents the results of analyzing the reasoning paths for questions that LMMs fail to answer correctly by *stan* prompting. The incorrect reasoning path could occur due to an incorrect spatial step, an incorrect non-spatial step, or a combination of both. On average, only 1% of questions had the correct reasoning path but an incorrect final answer. Moreover, 91% of questions included at least one incorrect spatial reasoning step, while only 9.75% contained at least one incorrect Non-spatial reasoning step. Please note that an incorrect reasoning step occurs when a reasoning step semantically does not match the ground truth reasoning steps or when a reasoning step from the ground truth path is missing in the generated path.

The average scores for all Spatial and Non-spatial reasoning steps in Spatial-CoT questions are listed in Table 4, which shows that the average gap between F1 scores for *Spatial* and *Non-spatial* reasoning steps is significant by 21 point on *stan* and 19 points on *stan+SG$_{synth}$* prompting.

Finding 3: Chain of thought prompt is not effective for multi-hop spatial reasoning.

Model	Gemini-Pro		LLaVA-1.5	
Data (GQA)	one_obj	two_obj	one_obj	two_obj
stan	19.40	16.35	27.16	15.12
stan+none	19.14	22.01	33.02	15.12
stan+none+bbox$_{gt}$	30.98	39.24	37.16	15.46
rel_neg+none	1.75	2.176	41.38	8.93
rel_neg+none+bbox$_{gt}$	0.43	0.35	16.98	2.41
rel_neg+none+SG	33.19	50.69	38.19	6.87
rel_swap+none	_	2.78	_	1.37
rel_swap+none+bbox$_{gt}$	_	2.07	_	0.00
rel_swap+none+SG$_{gt}$	_	64.58	_	0.00
obj_change+none	91.87	69.10	90.69	54.98
obj_change_a+none	45.76	30.21	61.55	52.58

Table 6: Results of different types of analyses on GQA-spatial dataset.

5.7 Analysis of Perturbations on GQA-spatial

Using GQA-spatial, we do a comprehensive understanding of the LMMs' robustness when facing spatial-related questions by applying **five perturbation settings** on the caption options as follows:

- none: adding one extra option "None of above"; the "Standard+None" row in Figure 4

- rel_neg: adding "not" to the correct option; the "Relation Changed (Negation)" row in Figure 4

- rel_swap: Swapping the *key objects* in options; "Relation Changed (Swapping)" row in Figure 4

- obj_change: replacing one of the key objects for both options (A & B); the "Object Changed (Two options)" row in Figure 4

- obj_change_a: replacing one of the key objects in the correct option (A) with a none-existing object in the image; the "Object Changed (option A)" row in Figure 4

For the Standard setting, the model input is an image paired with two caption options that differ only by the *preposition* they contain, from which LMMs should be able to select the caption option with the correct *preposition*. For perturbation settings, the image remains the same but caption options are changed with different difficulty levels. Intuitively speaking, we posit that all perturbation settings are more difficult than the Standard setting, as they add noise (e.g., none) and create obstacles (e.g., negation) for LMMs. Results are show in Table 6, and analysis of each perturbation are discussed in detail below:

rel_neg. Adding *not* to the answer option A increases the accuracy (7%) for 1-object only, meanwhile decreasing the accuracy (6%) for 2-objects. Negating the relations in general decreases model performance. Adding bounding boxes makes it

even worse. However, adding scene graphs is helpful, especially with Gemini.

rel_swap. Analysis is conducted with swapping objects of answer option A and adding the "None" option. From 288 two_obj questions, Gemini only has 8 correct answers (option E), 5 incorrect answers(option A), and 275 incorrect answers(option B), while 95% of generated answers are option B.

LLaVA shows similar results: from 291 2-object questions, LLaVA only has 4 correct answers (option E), 3 incorrect answers(option A), and 284 incorrect answers(option B), while 97.59% of generated answers are option B.

obj_change. When the same object in both options A and B are swapped with an object that does not exist in the image (row "obj_change+none" in Table 6), both Gemini and LLaVA show the highest accuracy scores in both one_obj and two_obj ((e.g. 91.87% and 69% respectively for Gemini). When only changing the object in the correct option (A), both models' performance decreases significantly. For instance, Gemini's one_obj performance decrease from 91.87% to 46% and two_obj performance from 69.10% to 30%.

> **Finding 4:** LMMs are usually good at the object detection task (recognizing objects present in the image), but struggle with spatial reasoning (distinguishing "left" from "right").

We also conducted an additional analysis of spatial propositions of *front-behind*, *left-right* and *top-bottom* in Appendix F. Interestingly, Gemini, GPT-4 vision and LLaVA show different performance characteristics across the propositions.

6 Conclusion

Motivated by our observation that LMMs are not able to distinguish spatial relationships, in this paper, we propose a new benchmark to evaluate the spatial reasoning capabilities of LMMs. Our benchmark consists of two subsets: Spatial-Obj containing 2,000 multiple-choice questions to evaluate the spatial reasoning capabilities of one or two objects in a given image, Spatial-CoT containing 310 multi-hop questions to evaluate the spatial related CoT and reasoning paths. We conduct comprehensive experiments on our benchmark and GQA-spatial. Experimental results show the deficiencies in spatial reasoning of current LMMs, including GPT-4o, Gemini and LLaVA-v1.5. We also find that bounding boxes and scene graphs are helpful in some cases to improve the LMMs' prediction quality.

7 Limitation

To set up the Spatial-CoT benchmark, we implement a two-step data collection process: initially using GPT-4 to generate multi-hop VQAs, followed by manual filtering and modification. Though this process can to some extend accelerate the overall annotation approach as human annotators don't need to write multi-hop VQAs from the scratch, it is still limited and hard to create and generate large scale of labelled data. Same limitations applies to the generation of the ground-truth reasoning paths.

Acknowledgments

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*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*.

Jihyung Kil, Farideh Tavazoee, Dongyeop Kang, and Joo-Kyung Kim. 2024. II-MMR: identifying and improving multi-modal multi-hop reasoning in visual question answering. *CoRR*, abs/2402.11058.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2024a. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2024b. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *Preprint*, arXiv:2402.12058.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023c. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023d. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023d. Visual instruction tuning. In *NeurIPS*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023e. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023f. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. 2023. Kosmos-2.5: A multimodal literate model. *CoRR*, abs/2309.11419.

Michael R Lyu, Baishakhi Ray, Abhik Roychoudhury, Shin Hwei Tan, and Patanamon Thongtanunam. 2024. Automatic programming: Large language models and beyond. *arXiv preprint arXiv:2405.02213*.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424.

Cristiane Kutianski Marchi Fagundes, Kristin Stock, and Luciene Stamato Delazari. 2021. A cross-linguistic study of spatial location descriptions in new zealand english and brazilian portuguese natural language. *Transactions in GIS*, 25(6):3159–3187.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In *ACL 2024*.

OpenAI. 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774.

OpenAI. 2024. Hello gpt-4o.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *CoRR*, abs/2306.14824.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2023. Rephrase, augment, reason: Visual grounding of questions for vision-language models. *arXiv preprint arXiv:2310.05861*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023a. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *CoRR*, abs/2310.02239.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *CoRR*, abs/2205.10625.

Jin Peng Zhou, Charles Staats, Wenda Li, Christian Szegedy, Kilian Q Weinberger, and Yuhuai Wu. 2024. Don't trust: Verify–grounding llm quantitative reasoning with autoformalization. In *ICLR 2024*.

A LMMs

In this paper, we evaluate the spatial reasoning capability of the following four LMMs: • **LLaVA-1.5-7B** (Liu et al., 2023b) is an open-source multimodal instruction-tuned model that achieves state-of-the-art performance on 11 benchmarks, with just simple modifications to the original LLaVA (Liu et al., 2023d).

• **GPT-4 Vision** (OpenAI, 2023) is a closed-source LMMs, which is an advanced capability of OpenAI's GPT-4 model. It is powerful to understand visual inputs such as images, and generate textual responses based on the content of images.

• **GPT-4o** (OpenAI, 2024), also known as GPT-4 Omni, is the latest and most advanced AI model. Improved upon GPT-4, it is able to handle multimodal inputs and outputs, including text and images, with the enhancement of multimodal capabilities and speedup over previous models.

• **Gemini 1.5 Pro** (Gemini Team, 2024) is a closed-source LMMs from Google Deepmind. Gemini 1.5 Pro is designed to be highly capable and general, excelling across various visual tasks with state-of-the-art performance.

• **MiniGPT-v2 (7B)** (Chen et al., 2023) is an improved version of MiniGPT, an open-source large multimodal model that enhances the ability to comprehend and describe visual information by combining a lightweight architecture with strong visual-language reasoning capabilities.

B Spatial-MM Benchmark Samples

The Figure 5 shows sampled instances from Spatial-MM benchmark.

C Pipeline

Figure 6 illustrates our pipelines for data enrichment.

D Bounding Box Example

Figure 7 shows examples of synthesized bounding boxes for the key objects along with the ground-truth bounding boxes.

E Spatial Relationships Types

Figure 8 demonstrates the distribution of spatial relationships in Spatial-MM.

Prepositions	All	front-behind	left-right	top-bottom
Number of instances	1451	26	1032	393
Gemini (stan)	18.52	15.38	19.16	17.05
Gemini (stan + bbox)	34.56	44.00	38.60	23.41
Gemini (stan + SG)	45.24	60.00	49.36	33.59
Gemini (stan + SG + bbox)	41.00	36.00	44.64	31.81
GPT-4v (stan)	23.71	79.17	18.18	63.10
GPT-4v (stan + bbox)	41.58	62.50	39.39	75.57
GPT-4v (stan + SG)	**77.32**	**95.83**	**75.38**	**100.00**
LLaVA (stan)	29.57	19.23	33.43	20.10
LLaVA (stan + bbox)	30.53	23.08	33.62	22.90
LLaVA (stan + SG)	41.83	46.15	48.93	22.90
LLaVA (stan + SG + bbox)	38.73	46.15	44.38	23.41

Table 7: Results categorized by different spatial prepositions on GQA-spatial dataset. We divided them by opposite preposition pairs: *front-behind*, *left-right*, and *top-bottom*.

F Preposition analysis

In Table 7, we present the model performance on different spatial relationships.

Gemini-Pro. For the *left/right* relations, including bounding boxes increases the accuracy(16-18%). However, by including SG, the accuracy increases significantly(26-37%)

For the *bottom/top* relations, while adding bounding boxes of the objects increases the accuracy slightly(one-obj only)(5%), adding SG increases the accuracy for one-obj questions by 15%.

For the *behind/front* relations, additional bounding box information does not change the accuracy or make it worse. however, including SG increases the accuracy significantly(for two-obj questions)(42%).

GPT-4-Vision. For the *left/right* relations, adding bounding boxes and SG increases the accuracy by 29% and 65%, respectively. Regarding *bottom/top* relations, while adding bounding boxes does not change the accuracy, adding SG increases the accuracy. For the *behind/front*, while including the bounding box, drops the accuracy by 13%, adding SG increases the accuracy by 28%.

> **Additional Finding**: The three LMMs exhibit different performance characteristics across the three spatial propositions.

G Prompts

Multi-hop QA Generation Prompt

Task: Given an image, please act as a linguistic master and write a complex multi-hop question that includes a spatial relationship (such as left, top, touching, behind, middle, front, etc.). Then, provide the answer.
Rules:
- The question should be limited to only one sentence.
- The answer should be limited to one word or phrase.
Examples:
<In context learning>

With these examples in mind, generate a multi-hop question including a spatial relationship between the objects and its answer for this <image>.
question:

Bounding Boxes Generation Prompt

Task: Given an image and key objects, please generate a list of bounding boxes for ONLY key objects.
The bounding boxes, indicating the position of objects in the image, which are represented as $[x_{min}, y_{min}, x_{max}, y_{max}]$ with floating numbers ranging from 0 to 1. x_{min} and y_{min} are coordinates of the top-left corner, and x_{max} and y_{max} are coordinates of bottom-right corner of the bounding box.
Rules:
- Identify the key objects and provide the bounding boxes in the form of {object: $[x_{min}, y_{min}, x_{max}, y_{max}]$
- Include object names.
- Provide accurate and complete coordinates for each bounding box.
- The number of bounding boxes should be less than 10.

Here is an example:
<In context learning>
————
With these examples in mind, please help me extract the bounding boxes of the given key objects in this <image>.
Input key objects: [object1, [object2, …]
Output bounding boxes:

Scene Graph Generation Prompt

Task: Given the image and its caption and key objects, please act as a linguistic master and extract a set of words describing the location, orientation, directions and spatial or positional relations between key objects in the image. Your answer should be a list of values that are in format of object1 relation with object2 with the relation being left, right, top, facing left, facing right, touching, behind, middle etc.
Rules:
- Do not extract any relation that is a verb, replace it with simply which object is (touching or in front of or etc) the other object or the image.
- Furnish the extraction enclosed within square brackets. If there are no relations found, please return an empty list.
Examples:
<In context learning>
————
With these examples in mind, extract the relation based on the information of the caption and key objects in this <image>.
caption: [caption]
Input key objects: [object1, object2, …]
Relations: :

Prompts for QA with Scene Graph/Bounding boxes

Given the scene graphs, <SCENE Graphs>, and this <image>, answer this <question> with a single word or phrase.

Given the bounding boxes, <BOUNDING BOXES>, and this <image>, answer this <question> with a single word or phrase.

Visual Pattern Classification Prompt

Task: Please act as a linguistic expert and review the questions and options, aiming to identify common patterns that vision-language models find challenging. Emphasize the visual features and generalize patterns that are crucial for vision-language models.

[question], [answer]
Visual pattern:

Is the dog that is lying atop the table facing towards the window?
A. no
B. yes

B. yes

Is there a car under the tree?
A. yes
B. no

A. yes

Which direction is this truck facing from the viewer's perspective?
A. left
B. right
C. front
D. back

A. left

Is this person using his right hand to hold the fork?
A. yes
B. no

B. no

Figure 5: Instances are identified where the spatial reasoning capabilities of GPT-4V (Achiam et al., 2023) fall short (Date accessed: June 6, 2024) due to inaccurate spatial understanding. Text in red signifies an incorrect response. All the images referenced are from our Spatial-MM benchmark which are sourced from Internet.

Question: where is the man holding a bowl in relation to the woman with purple hair seated?

Key objects Extraction

Object1: man holding a bowl
Object2: woman with purple hair

Bounding boxes : [
{man holding a bowl: [0.39, 0.25, 0.62, 0.65},
-{woman with purple hair: [0.23, 0.10, 0.40, 0.65}
]

Question: where is the man holding a bowl in relation to the woman with purple hair seated?

Key objects Extraction

Object1: man holding a bowl
Object2: woman with purple hair

Spatial-aware caption

Relations: [
{woman with purple hair to the left of man holding a bowl},
{man holding a bowl to the right of woman with purple hair}
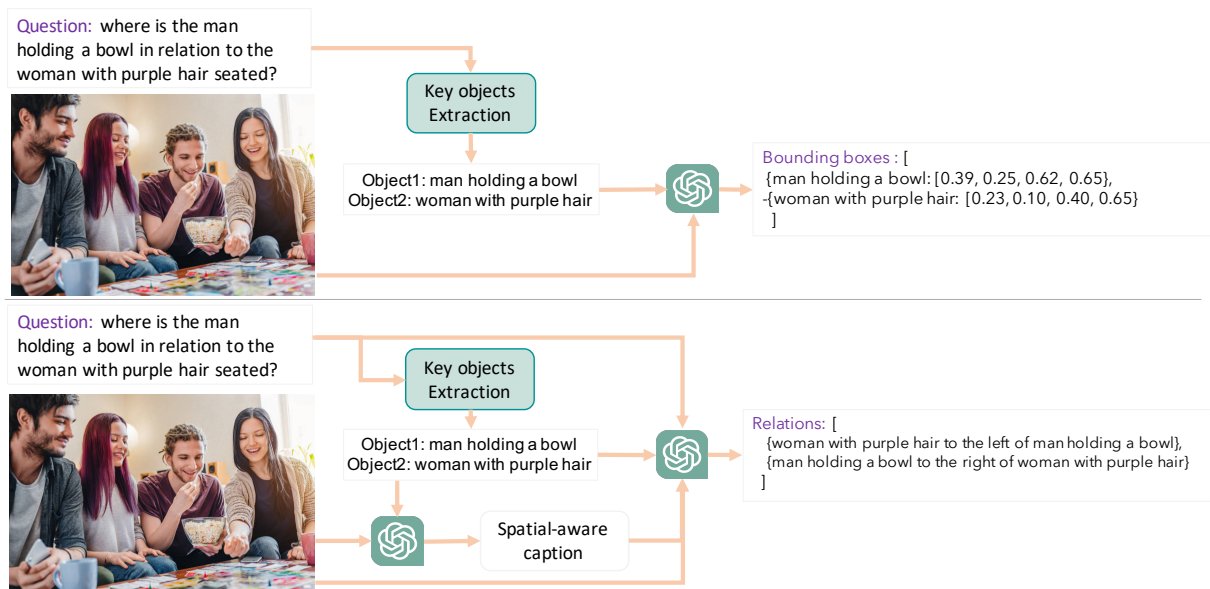]

Figure 6: The proposed pipelines for Bounding boxes generation (top) and Scene graph generation (bottom).

Figure 7: A comparison between the ground truth bounding boxes from GQA dataset on the left, and the synthesized bounding boxes by GPT-4o on the right.
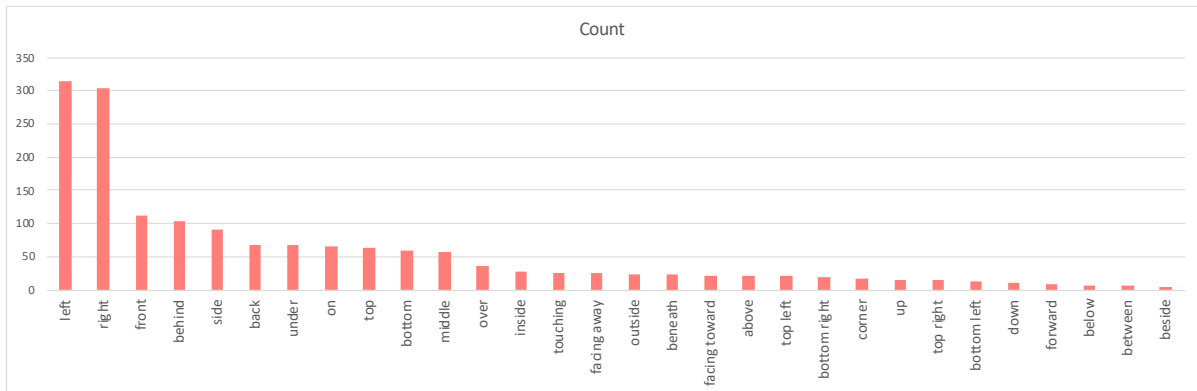
Figure 8: Relation distribution of Spatial-MM dataset(sorted by frequency). Top 30 most frequent relations are included.