See Beyond: Benchmarking MLLMs' Visual Relational Reasoning Ability

Yifan Wang[1] and Haizhou Wang[1(✉)]

[1] School of Cyber Science and Engineering, Sichuan University, Chengdu, China
whzh.nc@scu.edu.cn, sanqin@stu.scu.edu.cn

Abstract. Multimodal Large Language Models (MLLMs) have achieved remarkable success in visual and textual tasks. However, their visual relational reasoning capabilities remain insufficiently explored. Further more, most real-world Internet images used in model training lead models to guess answers via latent bias before actual reasoning. This highlights the need for further investigation into how MLLMs reason about relationships in abstract scenes. Recent studies suggest that visual relational reasoning is closely tied to intelligence, inspired by this, we introduce VRR-BENCH, a benchmark designed to evaluate the **V**isual **R**elational **R**easoning abilities of MLLMs. The dataset is divided into Non-Relational and Visual Relational reasoning tasks across three levels, each involving different combinations of object attributes. Our evaluation of six MLLMs, including GPT-4o, reveals that relational tasks are generally more challenging, with an average accuracy drop of 22.01%. Additionally, model performance varied with different numbers of object attribute combinations, indicating diverse challenges across tasks. We conducted comprehensive and progressive testing using VRR-BENCH, and we believe this research can serve as a reference for future work.

Keywords: Visual Relational Reasoning · MLLM · Benchmark

1 Introduction

"Simplicity is the ultimate sophistication." — Leonardo da Vinci

Multimodal Large Language Models (MLLMs) have demonstrated significant success in handling both visual and textual tasks. The visual spatial reasoning capabilities [1, 2] of these models are particularly crucial in fields such as robot navigation, embodied AI, or human-assistive systems [3–5] where the ability to understand and interpret the physical world is essential for intelligent decision-making.

Relational Tasks in Visual Spatial Reasoning. Recently, there has been many research on the visual spatial reasoning capabilities of MLLMs [6–8]. However, we observe a lack of studies regarding the reasoning performance of MLLMs when relational objects are involved, so-called visual relational reasoning. In the few existing studies, it is suggested that the ability to reason about the

relationships between entities and their attributes is central to general intelligent behavior [9].

Moreover, research also indicated that due to the majority of our real - world images on the Internet have been used in the training of various models, they often use latent bias to guess the answers before acutally reasoning. An example is: When asked "What is below the black-and-white object?" in a football scene, models may output a number matching "Field Zone" (relying on training biases) rather than performing real relation reasoning. Thus, to minimize such interference and focus more precisely on the models' reasoning processes, it is natural to chose generated abstract simple geometric objects to explore the performance of MLLMs in tasks involving relational objects. Fig. 1 includes a illustrative example: the non-relational question concentrates on the attributes of an individual object, while the relational question demands explicit reasoning regarding the relationships between objects, each corresponding to different problem texts derived from the same original image.

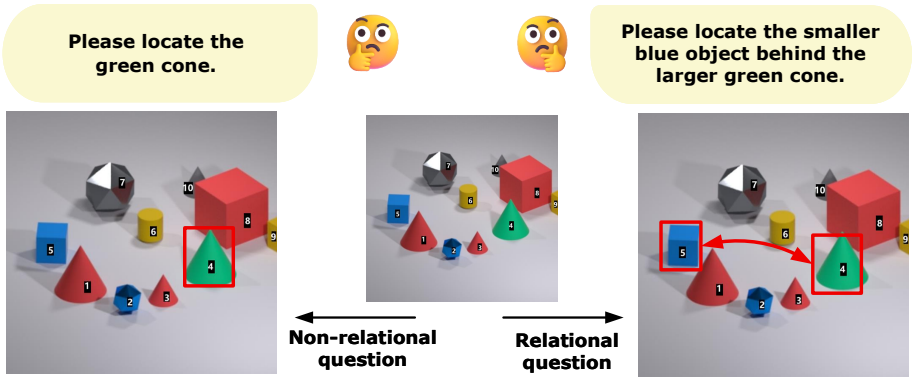


Fig. 1. Illustrative examples of Non-relation and Relation . **Left:** Non-Relational question concentrates on the attributes of a single object. **Right:** Relational reasoning question focuses on relationships among objects.

Research Questions. Based on the above discussion, we aim to explore the following research questions:

1. Does the involvement of relational objects affect the model's reasoning?
2. How different combinations of object attributes affect the model's final performance?

In response to these two questions, we have proposed **VRR-BENCH**[1] , a new benchmark designed to evaluate the **V**isual **R**elational **R**easoning abilities of MLLMs. The initial dataset was collected from an interactive CAPTCHA

[1] The whole dataset and code are now publicly available at https://github.com/yiyepianzhounc/VRR-BENCH.

(Completely Automated Public Turing Test to Tell Humans Apart), provided by Geetest[2] . We chose CAPTCHA images because they are designed to test human-like reasoning capabilities and minimize reliance on pre-trained biases, making them ideal for evaluating visual relational reasoning. Each image contains 7–10 objects, and every image–question pair is associated with a unambiguous answer, which is fully consistent with our research setting. As a benchmark emphasizing relational tasks, the dataset is evenly divided into two parts: one focusing on Non-Relational tasks and the other on Relational tasks. Each subset is categorized into three levels based on the number of attributes involved: Level 1 for a single attribute, and Level 3 for three attributes like color, shape, and size. Details can be found in Fig. 2. To be more specific, subsets are further divided into seven

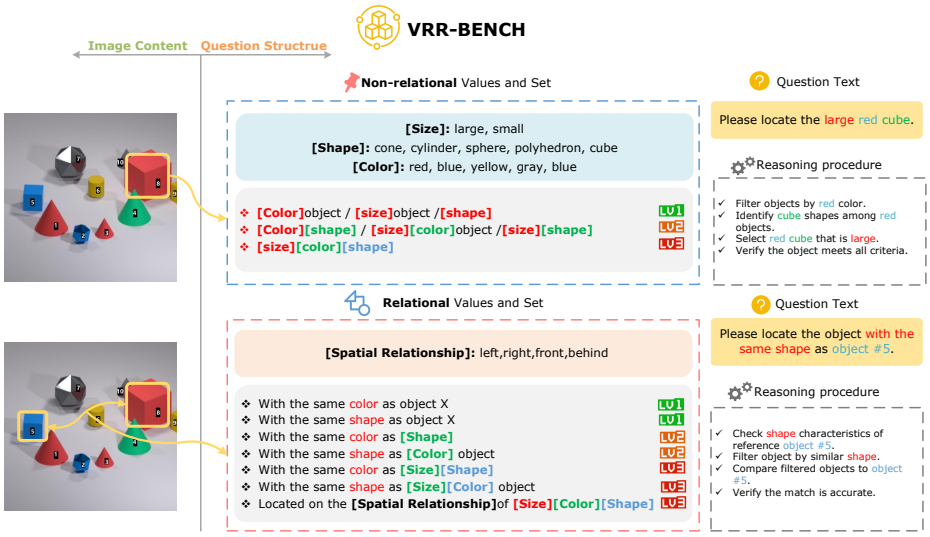


Fig. 2. The structure of VRR-BENCH.

categories, with each category comprising 100 image-question pairs, culminating in a total of 460 pairs. Subsequently, five experts were employed to revise the question texts according to our predefined classification of Non-Relational and Relational tasks. The dataset was further refined by adding object serial numbers using SoM (set-of-mark) [10], making our evaluation more objective.

With the complete dataset, we can examine the performance differences of the models in the presence or absence of relational objects, as well as their reasoning abilities across different attribute combinations. Detailed descriptions can be found in Section 3.

MLLMs Performance Evaluation. We conducted a thorough evaluation of the visual relational reasoning capabilities of six mainstream MLLMs using

[2] https://www.geetest.com/en/demo.

our VRR-BENCH dataset, including open-source models, such as InternVL-2-pro [11] and proprietary models like GPT-4o [12]. Our experiments and analysis lead to the following conclusions:

1. Relational tasks are generally more difficult than non-relational tasks, with an average performance drop of approximately 22.01%, and GLM-4V [13] showed the largest gap of 39.25%. Details in Section 4.
2. Further results, obtained by modifying the question texts of non-relational tasks to relational ones while keeping the images the same, confirm that relational objects notably affect the model's reasoning accuracy, with a 13.09% decline. This is demonstrated in more detail in Section 4.
3. The diversity and combination of object attributes variously affect the model's reasoning ability, as evidenced by additional tests revealing the model's difficulty in occlusion and size comparison tasks, with accuracies of only 36.67% and 38.33%, respectively. More details are provided in Section 5.

2 Related Work

2.1 Inspiration from Visual Spatial Reasoning

Early work such as CLEVR [2] created controlled 3-D scenes for probing spatial reasoning. SUPER-CLEVR [7] added object diversity, and Liu et al. [1] catalogued 66 relations. Yet many later datasets reuse Internet photos already seen during pre-training, letting models exploit bias instead of reasoning. For example, asked "What is below the black-and-white object?" in a football image, a model may output a field-zone label by association. To block such shortcuts we switch to generated scenes of simple geometric objects.

2.2 Dive into Relational Reasoning

The relational module of DeepMind [9], showed that explicit relations underpin intelligent behaviour,which motivated us to explore relational tasks more deeply. Building on this foundation, we probe how explicit relational objects shape model behaviour and clarify their contribution. We further posit that many systems may still struggle because visual simplicity does not equate to reasoning simplicity.

3 VRR-BENCH

Our process of constructing the benchmark can be divided into three stages:

3.1 Images and Question Collection

As a benchmark for visual relational reasoning, our images and questions are designed to cover a diverse range of relationships. We collected interactive CAPTCHA

images via the Geetest interactive API. All images contain between 7 and 10 objects, ensuring that the tasks are neither too simple nor too complex. In total, we selected 460 image-question pairs with a moderate number of objects for further analysis. This controlled scale ensures high annotation quality and task focus, and each CAPTCHA sample has been rigorously vetted to guarantee consistency and representativeness.

3.2 Fine-tune and categorize

To establish a systematic and progressively challenging benchmark, we have sorted all image-question pairs into two primary categories: ***Non-Relational*** tasks and ***Relational*** tasks. The former only involves reasoning related to attributes, while the latter introduces a relational task (A relational task involves comparing attributes and understanding spatial relationships between objects), which requires the model to first deduce the relational object before arriving at the correct answer.

Following this, we have organized them into three levels based on the attributes mentioned in the original question texts: color, shape, and size. (In our dataset, there are five colors, two sizes, and five shapes.) The complexity increases with each level, depending on whether the questions involve one or more of these attributes.

To further ensure the benchmark's quality, we had five annotators conduct another round of checks on all image-question pairs to eliminate any duplicates.

In summary, our benchmark's structure is depicted in Fig. 2. The left side represents images for the two different tasks, the center presents the basic statements of the question texts corresponding to these tasks (The basic statement is the most central part of a question text), and on the right, we offer examples along with the fundamental steps that models will follow during the reasoning process.

3.3 Add Serial Numbers

Due to the inherent unpredictability in model-generated responses, this leads to uncertainties in evaluation. Therefore, we employed SoM (Set-of-Mark) [10], a visual auxiliary processing method, to assign a unique number to each object in the image. This approach standardizes all model responses and meets the requirements for objective assessment. After all images were processed, our annotators conducted a second round to check the results.

4 Experiments

4.1 Experimental Setup

Models. Our study evaluated six MLLMs on VRR-BENCH, including three proprietary MLLMs (GPT-4o [12], Gemini-1.5-flash [14] and Qwen-VL-Max [15]

) and three open-source MLLMs (Moonshot-v1 [16], InternVL2-Pro [11] and GLM-4V [13]). Except Moonshot-v1, which lacks API file upload support, was tested via the web interface, other models were accessed through API calls.

Metrics. We use accuracy to assess model performance, which is the ratio of correctly answered questions to the total number of questions. A higher value indicates better model performance in delivering accurate responses.

Table 1. Overview of VRR-BENCH.

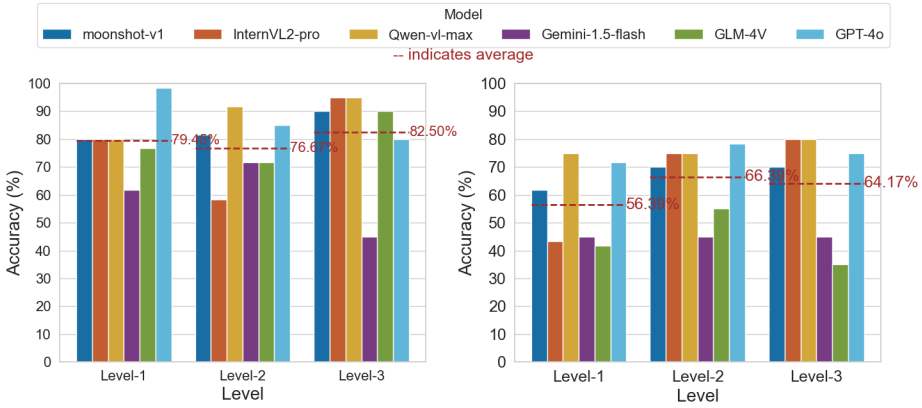
	Non-Relational	Relational	Average	Drop[a]
GPT-4o [12]	**90.00%**	75.00%	**82.50%**	16.67%
Qwen-VL-Max [15]	87.14%	**75.71%**	81.43%	**13.11%**
Moonshot-v1 [16]	82.14%	66.43%	74.29%	19.13%
InternVL2-Pro [11]	72.86%	62.14%	67.50%	14.71%
GLM-4V [13]	76.43%	46.43%	61.43%	39.25%
Gemini-1.5-flash [14]	63.57%	45.00%	54.29%	29.21%

All values in the table indicate accuracy.
Bold: *Best results.* *Underline: Second best results.*
[a]Drop represents the accuracy drop from the Non-Relational task to the Relational task.

4.2 Main Results

We compare the overall performance of different MLLMs in Table 1, reporting their average accuracy and the accuracy drop from Non-Relational task to relational task (shown in the last two columns). We observed that GPT-4o achieved the highest overall performance, followed closely by Qwen-VL-Max. In contrast, Gemini-1.5-flash ranked the lowest, with a significant performance gap compared to other models. For the Non-Relational subset, GPT-4o delivered the best results, and most models achieved accuracy rates above 75%, indicating that MLLMs generally perform well in recognizing individual object attributes. However, in the Relational subset, all models exhibited a decline in accuracy. Qwen-VL-Max demonstrated the smallest drop, maintaining the best performance in relational reasoning. Conversely, Gemini-1.5-flash and GLM-4V experienced accuracy drops exceeding 25%, suggesting that these models struggle more with relational reasoning tasks compared to attribute recognition.

Impact of Task Levels The two subplots in Fig. 3 present the average performance of different models on tests of varying levels. In Fig. 3(a), GPT-4o achieves the best performance at Level-1, while Qwen-VL-Max excels at both Level-2 and Level-3. Notably, Gemini consistently falls below the average performance across all tasks. In Fig. 3(b), Qwen-VL-Max outperforms all models across all tasks, while Gemini still lags behind the average. These results highlight that different models exhibit distinct strengths and weaknesses across tasks of varying difficulty.



(a) Performance in Non-Relational Task **(b)** Performance in Relational Task

Fig. 3. The average performance of each model across varying difficulty levels for Non-Relational tasks is depicted in Plot a, while their performance on relational tasks is shown in Plot b.

Performance on Sub-tasks Table 2 provides a detailed breakdown of model performance across various sub-tasks at different levels for both Non-Relational and Relational tasks. The sub-task name "1-shape" indicates that the task involves only the attribute of shape, with other sub-task names similarly referring to specific attributes, such as color or size.

In the Non-Relational task (upper half of Table 2), models generally perform well on color and shape recognition across all levels. For example, at Level-1, both Moonshot-v1 and GPT-4o achieve perfect accuracy on the 1-color sub-task. However, accuracy on size-related tasks is notably lower, with performance on the 1-size sub-task declining by nearly 20% compared to other single-attribute tasks like color and shape. Interestingly, we observe a surprising fact: among all models, Gemini-1.5-flash shows the worst performance on this sub-task, with an accuracy of only 15.00%. This trend suggests that in Non-Relational reasoning, size-related information is more challenging for models to process compared to attributes like color and shape.

For the Relational task (lower half of Table 2), the challenge of size-related reasoning remains prominent, with the average accuracy for the 1-size sub-task dropping to 48.33%. Notably, Gemini-1.5-flash performs the worst, with its accuracy falling to just 20.00%. These results suggest that relational reasoning, especially when dealing with size-related information, significantly increases the complexity of the task. Additionally, we observe that adding attributes without referents does not significantly challenge model's abilities.

Conclusion 1: Relational tasks are more challenging than Non-Relational ones, both task types perform poorly in size-related sub-tasks.

Table 2. The accuracy of each model on Non-Relational and Relational Sub-Tasks

Model	Non-Relational Level 1			Non-Relational Level 2			Non-Relational Level 3
	1-color	1-size	1-shape	2-color_size	2-color_shape	2-size_shape	3-color_size_shape
Moonshot-v1 [16]	**100.00%**	70.00%	70.00%	**95.00%**	70.00%	80.00%	90.00%
InternVL2-Pro [11]	80.00%	60.00%	**100.00%**	40.00%	**90.00%**	45.00%	**95.00%**
Qwen-VL-Max [15]	70.00%	85.00%	85.00%	**95.00%**	**90.00%**	90.00%	**95.00%**
Gemini-1.5-flash [14]	95.00%	15.00%	75.00%	70.00%	65.00%	80.00%	45.00%
GLM-4V [13]	75.00%	70.00%	85.00%	70.00%	70.00%	75.00%	90.00%
GPT-4o [12]	**100.00%**	**95.00%**	**100.00%**	75.00%	85.00%	**95.00%**	80.00%
average	86.67%	65.83%	85.83%	74.17%	78.33%	77.50%	82.50%
Model	**Relational Level 1**			**Relational Level 2**			**Relational Level 3**
	1-color	1-size	1-shape	2-color_size	2-color_shape	2-size_shape	3-color_size_shape
Moonshot-v1 [16]	55.00%	55.00%	75.00%	80.00%	45.00%	**85.00%**	70.00%
InternVL2-Pro [11]	45.00%	**65.00%**	20.00%	**85.00%**	**85.00%**	55.00%	**80.00%**
Qwen-VL-Max [15]	**90.00%**	55.00%	**80.00%**	80.00%	65.00%	80.00%	**80.00%**
Gemini-1.5-flash [14]	35.00%	20.00%	**80.00%**	60.00%	20.00%	55.00%	45.00%
GLM-4V [13]	40.00%	30.00%	55.00%	65.00%	50.00%	50.00%	35.00%
GPT-4o [12]	70.00%	**65.00%**	**80.00%**	80.00%	**85.00%**	70.00%	75.00%
average	55.83%	48.33%	65.00%	75.00%	58.33%	65.83%	64.17%

A comparative overview of six models' accuracy across Non-Relational and Relational sub-tasks. ***Bold****: Best results.* *Underline: Second best results.*

4.3 The Effect of labeling objects

As mentioned before, the images in our dataset were captured from the visual inference CAPTCHA, and there was no auxiliary information in the original images. The necessity of using auxiliary information, such as the addition of visual information through SoM [10], can be explained in the following two aspects:

(1) **As for model inference**, they need to give the geometrical center coordinates of the correct object. Even if a scale range is specified, MLLM will still output answers that are not within the correct range, as shown in Table 3, the performance of each model drops dramatically when the auxiliary ordinal is not used, however, this does not equate to a decrease in reasoning ability.

(2) **As for model performance evaluation**, the coordinates output by each model are a vague range, which makes it impossible to objectively determine how far the output answers deviate from the geometric centre of the correct object during evaluation. Furthermore, since our study emphasizes the reasoning ability concerning object identification, the models only need to accurately point to the correct object, thus finer granularity in the outputs is not required.

4.4 Internal Comparisons for Validation

The analysis in Section 4.2, reveals that the model performs better on the non-relational task compared to the relational task. This performance gap is likely due to the additional complexity of the latter, which involves relational objects. Specifically, the relational task requires the model to first identify the relational object to arrive at the correct answer. However, as the test data for the two tasks

Table 3. Model Comparison with or without Set-of-Mark for Labelled Serial Numbers

Model	w/ SoM[a]	w/o SoM
Moonshot-v1	70.00%	20.00%
InternVL2-Pro	**80.00%**	25.00%
Qwen-VL-Max	**80.00%**	**35.00%**
Gemini-1.5-flash	45.00%	25.00%
GLM-4V	35.00%	20.00%
GPT-4o	75.00%	20.00%
Average	64.17%	24.17%

[a] *w: with w/o: without*

Table 4. Accuracy Comparison of Different Models on Occlusion and Size comparison tasks

Model	Occlusion	Size
Moonshot-v1	35.00%	25.00%
InternVL2-Pro	**65.00%**	50.00%
Qwen-VL-Max	25.00%	**60.00%**
Gemini-1.5-flash	10.00%	25.00%
GLM-4V	30.00%	15.00%
GPT-4o	55.00%	55.00%
Average	36.67%	38.33%

All values in the table indicate accuracy. ***Bold:*** *Best results.* Underline: Second best results.

differ in structure and design, the observed disparity may also partially result from variations in the data distribution.

To investigate whether the Relational subset is an effective design for evaluating spatial reasoning with relational objects, We introduced a new data subset called the Aux-Relational Task, which is the Auxiliary subset. All images in this subset are the same as those in the Non-Relational task, but the problem text has been completely redesigned to include relational objects.

This contrast underscores the unique challenge posed by the Relational subset: the need for the model to process and integrate relational relationships to infer the solution. It also highlights the design rationale of *Aux-Relational* task as a bridge between *Non-Relational* and *Relational* task, allowing for a more nuanced evaluation of the model's relational reasoning capabilities.

Subsequently, we compared the model's performance across the Non-Relational, Aux-Relational, and Relational tasks.

As shown in Fig. 4, although the model used relational objects in the aux-relational task, its reliance on them was less pronounced than in the relational task. Specifically, the model's accuracy dropped by 13.09% when moving from the Non-relational to the Aux-Relational task. This proves that relational objects indeed increase the model's reasoning difficulty, demonstrating the effectiveness of the Relational subset.

Conclusion 2: Simplicity in scenes does not equate to simplicity in reasoning: Relational objects increase reasoning difficulty.

5 Further Analysis

In previous experiment, we observed that the model performed poorly when handling cases where the correct object was occluded. Additionally, the performance

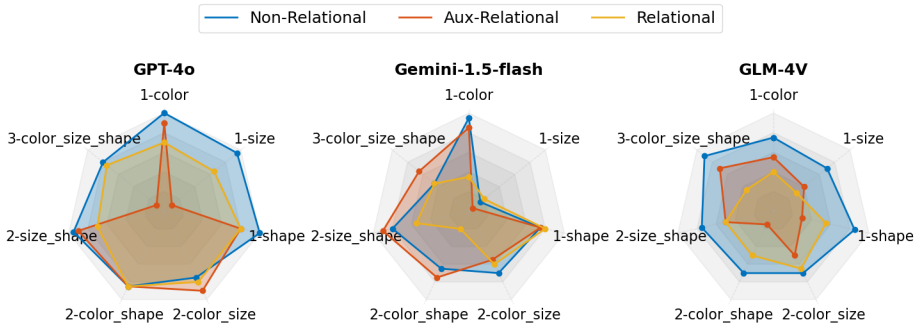


Fig. 4. A decline in performance is observed across three representative models as the transition from ***Non-Relational*** task, through ***Aux-Relational*** task, to ***Relational*** task, indicating that the incorporation of relational objects indeed demands greater visual relational reasoning capabilities from the models.

of all models on tasks involving size attributes was significantly lower than on tasks involving other attribute combinations. These observations prompted us to conduct further investigations.

5.1 The Influence of Occlusion Object

To build on our earlier overview, we now turn to a more detailed analysis of specific challenges observed in the experiments.

In the new occluded data subset, where all correct objects are partially obscured, most models experienced a sharp decline in accuracy, as shown in the "Occlusion" column of Table 4. InternVL2-Pro performed the best, maintaining an accuracy of 65%. These results suggest that recognizing partially obscured objects is inherently challenging, likely due to the models' lack of a systematic global observation mechanism for object identification.

5.2 Size Comparison is More Difficult

The size comparison subset requires identifying the largest or smallest object first, then determining the relative sizes of the remaining objects. Despite this task involving repeated reasoning about object sizes, all models still showed a significant drop in performance. As shown in the "Size Comparison" column of Table 4, Qwen-VL-Max achieved the highest accuracy at only 60%. This suggests that the models have room for improvement in long-term memory, particularly in recalling previously excluded objects, which is essential for accurate reasoning.

6 Conclusion

In this paper, we presented VRR-BENCH, a benchmark dataset for studying the visual relational reasoning capabilities of models, including two subsets: Non-

Relational reasoning for individual objects and Relational reasoning for involving referents, and conducted comprehensive experiments based on this benchmark. The models we tested include six widely-used models, both from open-source and proprietary sources, and our results show that major MLLMs, including GPT-4o, are still lacking in visual relational reasoning tasks. We conducted an in-depth analysis of the effectiveness of the relational subset by introducing an bridge data subset: the Aux-Relational subset. In addition to this, we found in our further study that the model especially needs to improve its reasoning ability regarding occluded objects and size comparison tasks.

7 Acknowledgement

This work is supported by the National Key Research and Development Program of China under grant No. 2022YFC3303101.

References

References

1. F. Liu, G. Emerson, and N. Collier. Visual spatial reasoning. Transactions of the Association for Computational Linguistics, 11:635–651, 2023.
2. J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2901–2910, 2017.
3. H. Le, S. Saeedvand, and C. Hsu. A comprehensive review of mobile robot navigation using deep reinforcement learning algorithms in crowded environments. Journal of Intelligent & Robotic Systems, 110:158, 2024.
4. A. Brohan. Rt-2: Vision–language–action models transfer web knowledge to robotic control. In Proceedings of the Conference on Robot Learning, pages 2165–2183, 2023.
5. A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich, F. Xia, C. Finn, and K. Hausman. Open-world object manipulation using pre-trained vision–language models. In Proceedings of the Conference on Robot Learning, pages 3397–3417, 2023.
6. R. Wadhawan, H. Bansal, K. Chang, and N. Peng. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. arXiv preprint arXiv:2401.13311, 2024.
7. Z. Li, X. Wang, E. Stengel-Eskin, A. Kortylewski, W. Ma, B. Van Durme, and A. Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14963–14973, 2023.
8. B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia. 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, 2024.

9. A. Santoro, D. Raposo, D. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In Proceedings of the Annual Conference on Neural Information Processing Systems, pages 4967–4976, 2017.
10. J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441, 2023.
11. Z. Chen. Internvl: Scaling up vision foundation models and aligning for generic visual–linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 24185–24198, 2024.
12. OpenAI, :, Aaron Hurst, and Adam Lerer et.al. Gpt-4o system card, 2024.
13. A. Zeng. Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793, 2024.
14. M. Reid. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.
15. J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision–language model for understanding, localization, text reading, and beyond. arXiv preprint, 2023.
16. R. Qin, Z. Li, W. He, M. Zhang, Y. Wu, W. Zheng, and X. Xu. Mooncake: A kvcache-centric disaggregated architecture for llm serving. arXiv preprint arXiv:2407.00079, 2024.
17. D. Hudson and C. Manning. 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, 2019.
18. S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In Proceedings of the Conference on Robot Learning, pages 892–909, 2022.
19. N. Rajabi and J. Kosecka. Towards grounded visual spatial reasoning in multimodal vision language models. arXiv preprint arXiv:2308.09778, 2023.
20. K. Wei, Y. Fu, Y. Zheng, and J. Yang. Physics-based noise modeling for extreme low-light photography. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44:8520–8537, 2021.
21. N. Methani, P. Ganguly, M. Khapra, and P. Kumar. Plotqa: Reasoning over scientific plots. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1527–1536, 2020.
22. Y. Li, B. Hu, H. Shi, W. Wang, L. Wang, and M. Zhang. Visiongraph: Leveraging large multimodal models for graph theory problems in visual context. arXiv preprint arXiv:2405.04950, 2024.
23. Zhengxuan Zhang, Yin Wu, Yuyu Luo, and Nan Tang. Fine-grained retrieval-augmented generation for visual question answering. arXiv preprint arXiv:2502.20964, 2025.
24. Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 5481–5489, 2024.
25. Yu Cheng, Arushi Goel, and Hakan Bilen. Visually interpretable subtask reasoning for visual question answering. arXiv preprint arXiv:2505.08084, 2025.
26. Xiang Shen, Dong Han, Le Zong, Zhi Guo, and Jun Hua. Relational reasoning and adaptive fusion for visual question answering. Applied Intelligence, 54(6):5062–5080, 2024.