

Appendices

Contents

1	Introduction	1
2	Related Works	3
3	MaRs-VQA Dataset	4
4	Problem Statement	4
5	Methods	5
5.1	Multi-Image Reasoning via Chain-of-Thought (CoT)	5
5.2	SFT for Vision-Language Model (VLM)	6
6	Experiments	6
6.1	Experimental Settings	6
6.2	Experimental Results	7
6.3	Ablation Study	8
6.4	Visualization	9
7	Discussion	10
8	Conclusion	10
	Appendices	16
A	Data Collection and Licenses	17
B	MaRs-VQA Difficulty Level Study	17
C	Experimental Settings	18
C.1	Implementation Details	18
C.2	More Qualitative analysis	19
D	Further Discussion on Limitations	19
E	Ethics Discussion	21
E.1	Negative Societal Impacts	21
E.2	Mitigating Bias and Negative Societal Impacts	21

A Data Collection and Licenses

We showed and compared MaRs-VQA and RAVEN in Table 9. The reason we choose RAVEN, MaRs-VQA is because all these datasets contain zero-shot / few-shot human investigation results in their follow-up studies. Based on these results, we can compare the MLLM’s performance with human in matrix reasoning tasks.

For RAVEN, we followed the original data generation pipeline in their repo. For MaRs-VQA, we download all questionnaires from MaRs-IB and then re-annotate all images by ourselves.

RAVEN The original dataset link of RAVEN is github.com/WellyZhang/RAVEN. It is under GPL-3.0 License (RAVEN LICENSE) and is free to use by public. All data in RAVEN are generated by rule-based scripts. We follow the basic setting of RAVEN, and modify the range of COLOR_VALUES to [255, 192, 128, 64, 0] and SIZE_VALUES to [0.3, 0.45, 0.6, 0.75, 0.9]. The sample size of RAVEN is 560.

MaRs-VQA The image data of MaRs-VQA is from MaRs-IB (Chierchia et al., 2019) and annotated with context option by our team. It contains 18 questionnaires, each of questionnaire contains 80 matrix reasoning questions. The human study of MaRs-IB is rigorous. In MaRs-IB’s original user study, all participants provided informed consent and all procedures were approved by UCL’s ethical committee.

The paper and study results are under MIT License. All questionnaires are under Attribution-NonCommercial 3.0 (MaRs-IB LICENSE), which means it allows people to use the work, or adaptations of the work, for noncommercial purposes only, and only as long as they give credit to the creator. Thus, the MaRs-VQA dataset will under the same license.

The sub-task statistics of MaRs-VQA is in Table.

Compared to other zero-shot matrix reasoning dataset (Table 1) to evaluate matrix reasoning for MLLMs, MaRs-VQA has advantages list below:

- MaRs-VQA comprises 1,440 image instances designed by psychologists, making it the largest dataset for zero-shot matrix reasoning evaluation.
- MaRs-VQA includes a diverse range of data, such as variations in color, geometry, positional relationships, and counting.
- The data source for MaRs-VQA is MaRs-IB (Chierchia et al., 2019), which is based on rigorous human studies. This dataset is widely recognized in the psychology community and has inspired numerous follow-up studies in child psychology and pediatrics. This is the first time we introduce it to the AI/ML community.

B MaRs-VQA Difficulty Level Study

We also compare GPT-4o across difficulty levels and different visual complexities in the MaRs-VQA dataset in Table 8. In Table 7, the difficulty of matrix reasoning tasks can be categorized into five levels based on the complexity of attribute changes: Difficulty Level 1 involves a single changing attribute (e.g., shape, color, size, position, or multi-object) or two simple attributes; Difficulty Level 2 combines multi-object attributes with one other attribute (e.g., shape, color, size, or position); Difficulty Level 3 involves three simultaneously changing simple attributes (e.g., shape, color, and size); Difficulty Level 4 combines multi-object attributes with two other attributes (e.g., shape and color); and Difficulty Level 5 and above includes combinations of four or more attributes. The difficulty increases as the number and complexity of attribute combinations grow. The results indicate that GPT-4o exhibits difficulty sensitivity similar to that of humans. This is because GPT-4o can solve object size sub-tasks well in the MaRs-VQA, but is still struggling with other sub-tasks, especially the multi-object sub-task.











Difficulty Level	Question	Option	Description
Level 1			Shape + Size
Level 2			Color + Multi-object
Level 3			Shape + Color + Position
Level 4			Shape + Color + Multi-object
Level 5			Shape + Color + Position + Multi-object

Table 7: Explanation of Difficulty Levels in MaRs-VQA.

Method	Accuracy (%) \uparrow				
	Level 1 (90)	Level 2 (96)	Level 3 (84)	Level 4 (72)	Level 5 (138)
GPT-4o + CoT	57.78	27.08	27.38	19.43	21.74

Table 8: Test GPT-4o with different difficulty levels in MaRs-VQA. The number in the “()” is the number of case sample of selected level. The difficulty level is based on the complexity of color, size, geometry, positional relationships, and object counting (See Appendix for more details).

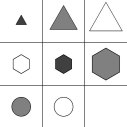
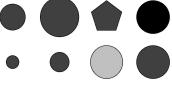


Dataset	Question	Option	Instance	Description
RAVEN (Zhang et al., 2019)			rule-based generation	8 options per instance grayscale image rule-based stimuli include human study
MaRs-VQA			1,440	4 options per instance RGB image psychologist designed stimuli include human study

Table 9: Experiment datasets comparison. The OOD dataset RAVEN is rule-based generated datasets. The test samples in MaRs-VQA are designed by psychologists from MaRs-IB.

C Experimental Settings

C.1 Implementation Details

We used langchain to implement all closed-source MLLMs. The temperature of all models are 0 and the max token length is 1024. For all datasets, we follow their default image size,

type settings for closed-source MLLMs. All experiments are run with three different random seeds, however, since we set temperature to 0, the final accuracy is the same for all random seeds.

For open-source models, we use the public available weights and data loader settings from the HuggingFace. Testing is conducted using two NVIDIA H100 GPUs for all VLMs. All experiments are run with three different random seeds, and the results are averaged.

Based on Figure 7, here is the explanation of difficulty levels presented in our paper:

- **Difficulty Level 1:** Single sub-task and two simple sub-tasks Description: The task involves only one changing attribute across the matrix reasoning—either shape, color, size, position, or multi-object. Or two simple attributes: (shape & color), (shape & size), (shape & position), (color & size), (color & position), (size & position). Example: Figure 4 (top-left) is a matrix reasoning task where only the size and color of the objects changes. This is a difficulty level 1 task.
- **Difficulty Level 2:** Two sub-tasks involving multi-object sub-task Description: The task involves multiple objects combined with one other changing attribute. The sub-task combinations are (multi-object & shape), (multi-object & color), (multi-object & size), (multi-object & position).
- **Difficulty Level 3:** Three simple sub-tasks combined Description: The task involves three changing attributes simultaneously. The sub-task combinations are (shape & color & size), (shape & position & size), (shape & position & color), (size & position & color).
- **Difficulty Level 4:** Three sub-tasks involving multi-object sub-task Description: The task involves multiple objects combined with two other changing attributes. The sub-task combinations are (multi-object & shape & color), (multi-object & shape & size), (multi-object & shape & position), (multi-object & color & position), (multi-object & color & size), (multi-object & position & size).
- **Difficulty Level 5 and Above:** Four or more Sub-tasks Description: The task involves combinations of four or five attributes. Example: Figure 4 (top-right) is a matrix reasoning task (shape & position & color & multi-objects) and its difficulty level is > 4.

As more attributes change simultaneously, the task becomes more complex, requiring higher levels of abstract reasoning to identify patterns. In addition, each additional changing element adds to the cognitive load, making it more challenging to discern the correct answer.

C.2 More Qualitative analysis

In this section, we further analyze the failure cases of GPT-4o. Correct reasoning is highlighted in green, while incorrect reasoning is marked in red. Although GPT-4o is sometimes able to extract a subset of key information from the question image, it frequently fails to arrive at the correct final answer. This is primarily due to critical features being either overlooked or inadequately utilized in the decision-making process. As a result, the final answers are often incorrect or only partially aligned with the relevant attributes. It reveals that visual working memory will be a key part to optimize the MLLM’s performance in matrix reasoning problem.

D Further Discussion on Limitations

Insights Unlike other VQA benchmarks, our work approaches the perspective of human visual cognition—an underexplored domain. Based on our experimental results, we offer the following insights for vision researchers:

- While scaling laws have some applicability to visual cognition tasks, merely increasing model size and training data is insufficient to achieve human-level performance.

- To demonstrate that VLMs possess strong visual cognitive abilities, it is crucial to evaluate them on zero-shot inference tasks like matrix reasoning—tasks characterized by simple visual content but requiring complex reasoning to find the correct answer.
- Unlike other multi-image visual reasoning benchmarks, MaRs-VQA effectively highlights the performance gap between MLLMs and human cognition in these tasks.

From our main and ablation experiments, we observed that as task difficulty increases, the performance of MLLMs in multi-image reasoning scenarios deteriorates. Interestingly, providing language-based descriptions of each option (i.e., inputting the model with a single question image and context-based options) improved the models’ performance compared to using multi-image options. This suggests that language still plays a significant role in the visual reasoning processes of current MLLMs and VLMs.

In contrast, human visual cognition—especially in children—allows individuals to solve matrix reasoning tasks without relying on advanced language reasoning capabilities. Children can often solve these tasks effectively by utilizing their visual working memory and pattern recognition skills.

One potential reason for the performance gap is that current MLLMs/VLMs may under-emphasize the visual encoder relative to the language encoder. In many recently released VLMs, the visual module is much smaller than the language model module, and the visual encoders are frozen during Large Language Model (LLM) and alignment layer fine-tuning in open-sourced VLMs. This imbalance might limit the models’ capacity to retain and process complex visual information during reasoning tasks.

To better retain visual information during the reasoning process, MLLMs may require more capable visual modules that can handle complex visual patterns and maintain this information throughout the reasoning steps. Moreover, optimizing the training process with end-to-end multimodal training—without freezing any layers in the visual modules—can be beneficial. Recent models have begun to explore end-to-end VLM fine-tuning, demonstrating the potential of this approach, though challenges remain such as the need for multi-round alignment. In the future, developing more advanced methods to effectively integrate visual and linguistic features will be crucial.

Limitations In the main paper, we briefly discussed the limitations of our work. Here, we provide a more in-depth discussion. First, our dataset is composed of limited publicly available matrix reasoning datasets, which must include human study results. The RAVEN, created by the AI/ML community, were not developed following rigorous psychological research norms. Consequently, our benchmarking results, which utilize these datasets, should not be used to derive psychological or clinical conclusions. While MaRs-VQA addresses this problem, its samples cannot represent all formats of matrix reasoning found in IQ tests such as the WISC and the Cattell Culture Fair Intelligence Test (Cattell & Cattell, 1960). We cannot use these IQ tests directly because they are not freely available, and copyright restrictions usually prevent these pen-and-paper tasks from being adapted into computerized formats.

Second, the size of MaRs-VQA is relatively small compared with typical computer vision datasets, due to the inherent challenges involved in collecting matrix reasoning data. However, as we have argued in our paper, matrix reasoning should not be presented in typical machine learning settings—fine-tuning models on training sets and evaluating performance on test sets. Benchmarking MLLMs’ visual reasoning performance should be conducted in a zero-shot inference setting, ensuring that all data in the test set are not included in the models’ training data. Even compared with other recently released human-designed matrix reasoning datasets, ours is still the largest (see Table 1).

Future Work Finally, we pose the open-ended question of whether MLLMs need to achieve or surpass human-level zero-shot inference capability in matrix reasoning tasks. Addressing this issue requires drawing on theories from cognitive science and psychology to understand

the nature of human and MLLM intelligence. Matrix reasoning ability develops early in human neurodevelopment, with children as young as four providing sensible answers to simple matrix reasoning questions without additional training, making it a critical component of IQ tests. In contrast, LLMs and MLLMs rely on training data, fundamentally differing from how children develop cognitive abilities. However, we believe that these two learning processes share commonalities: both involve the gradual accumulation of skills and the ability to generalize from past experiences. Exploring these parallels can provide valuable insights into designing MLLMs that more closely mimic human visual cognition, ultimately leading to more advanced and capable models. Additionally, we observe that current open-source models achieve matrix reasoning performance very close to that of closed-source models. However, VLMs face challenges in supporting multiple images as input and managing visual memory. Addressing these challenges is a crucial direction for building more robust open-source VLMs in the future.

E Ethics Discussion

This research aims to advance LLMs and VLMs by providing a new benchmark for evaluating AI capabilities in visual reasoning. MaRs-VQA builds on the MaRs-IB (Attribution-NonCommercial 3.0 License), RAVEN (GPL-3.0 License). All code and data are available on GitHub. No conflicts of interest exist among the study’s contributors. The annotation process is IRB approved by a clinical institute.

E.1 Negative Societal Impacts

We foresee no direct negative societal impacts from our matrix reasoning benchmark. However, it could be misunderstood or misinterpreted as comparing AI “thought” to human cognition or misused to evaluate human abilities across demographics or ethnicity. We strongly caution against such misuse, as our datasets are not validated for human assessment.

Another concern relates to the future conclusion from our benchmark. While matrix reasoning is a crucial test for evaluating human intelligence, observing that VLMs with large model weights perform better on matrix reasoning tasks does not imply that the intelligence of MLLMs follows the same “scaling law” from the general domain. A comprehensive intelligence test requires accurate assessment using human-based tools, of which matrix reasoning is only one critical component. We cannot conclude that larger MLLMs can achieve human intelligence.

Additionally, there is a potential concern for discrimination against certain groups based on race, gender, or age in human study results. Although all human results in our experiment tables are sourced from previously published papers, we cannot guarantee that all previous research adhered to strict standards ensuring the inclusion of all groups in the human investigation process.

E.2 Mitigating Bias and Negative Societal Impacts

While the use of MaRs-VQA come with potential negative social impacts, there are viable mitigations that can address these concerns. These include adding instructions for proper use and restricting unethical human investigations. Users must be aware of the ethical implications associated with our benchmark and take appropriate measures to ensure its safe and responsible utilization.

Question Image	Options	Reasoning
		In each row, the top shape is the same, and the bottom two shapes are always pink. Following this pattern, the missing shape should have the same top shape as the others in its row, which is pink, with two pink shapes at the bottom.
		Each row contains the same three shapes in different positions. The missing image should include the red circle with a blue plus, a blue circle, and a red triangle. Option 3 matches this pattern.
		Each row has a pattern where the number of blue dots increases by one as you move from left to right. The first column has 4 dots, the second has 5, and the third has 6. Following this pattern, the missing shape should have 6 dots.
		In each row, the number of dots in the top part increases from left to right (1, 2, 3). Following this pattern, the bottom-right cell should have three dots on the top, matching the third column's pattern for the row.
		In each row, a circle and a pair of parallel lines appear, with the circle shifting positions and colors. The third column should have a pink circle and pink lines to complete the pattern in the last row.
		Each row follows a pattern where the main shapes alternate positions, with the half-circle and ellipses appearing in different orientations. The third row should mirror this pattern. In the first row, the half-circle is on the left, in the second row, it is on the right, and in the third row, it should be in the middle.

Figure 5: More visualization results for GPT-4o's reasoning.