BENCHMARKING VISUAL COGNITION OF MULTI-MODAL LLMS VIA MATRIX REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, Multimodal Large Language Models (MLLMs) and Vision Language Models (VLMs) have shown great promise in language-guided perceptual tasks such as recognition, segmentation, and object detection. However, their effectiveness in addressing visual cognition problems that require high-level multi-image reasoning and visual working memory is not well-established. One such challenge is matrix reasoning – the cognitive ability to discern relationships among patterns in a set of images and extrapolate to predict subsequent patterns. This skill is crucial during the early neurodevelopmental stages of children and is proven to be used to test human intelligence. Inspired by the matrix reasoning tasks in Raven's Progressive Matrices (RPM) and Wechsler Intelligence Scale for Children (WISC), we propose a new dataset MaRs-VQA and a new benchmark VCog-Bench to evaluate the zero-shot visual cognition capability of MLLMs and compare their performance with existing human intelligent investigation. Our comparative experiments with different open-source and closed-source MLLMs on the VCog-Bench revealed a gap between MLLMs and human intelligence, highlighting the visual cognitive limitations of current MLLMs. We believe that the public release of VCog-Bench, consisting of MaRs-VQA, and the inference pipeline will drive progress toward the next generation of MLLMs with human-like visual cognition abilities.

1 INTRODUCTION

Matrix reasoning is a crucial ability in human perception and cognition, essential for nonverbal, culture-reduced intelligence measurements as it can minimize the influence of acquired knowledge and skills (Jensen, 1998; Jaeggi et al., 2010; Laurence & Macedo, 2023). Common matrix reasoning problems consist of images with simple shapes governed by underlying abstract rules (Małkiński & Mańdziuk, 2023) (see Figure 1). Participants have to identify and comprehend the rules based on a few provided patterns, and then reason about the next pattern following the same rules. Matrix reasoning is an important reflection of many fundamental capabilities of human intelligence, such as processing speed and working memory, that emerge in the early stage of children's neurodevelopment (Gentner, 1977). To quantitatively measure human's intelligence using matrix reasoning, many assessment methods have been proposed as a part of fluid intelligence tests. The two most famous assessments are Wechsler Intelligence Scale for Children (WISC) (Wechsler & Kodama, 1949) and Raven's Progressive Matrices (RPM) (Raven, 2003).

Recently, matrix reasoning tasks have emerged as an ideal testbed for investigating whether deep learning models can match or even surpass human cognitive abilities, motivating the creation of diverse problem settings and datasets (Chollet, 2019; Małkiński & Mańdziuk, 2023; Barrett et al., 2018; Zhang et al., 2019; Webb et al., 2020). However, previous research on matrix reasoning assessments applied typical machine learning settings – finetuning models on training sets and evaluating the performance on test sets (Hu et al., 2021; Małkiński & Mańdziuk, 2022; Zhao et al., 2024). This setting makes current matrix reasoning assessment an ill-posed problem because such tests accurately reflect reasoning capability only when subjects engage *without* prior training, *i.e.*, in zero-shot inference settings. Thus, establishing a matrix reasoning benchmark tailored for deep learning models remains an unsolved problem. Recently, Multimodal Large Language Models (MLLMs) have shown surprising understanding and reasoning capabilities, marking an important milestone towards Artificial General Intelligence (AGI) (Chollet, 2019; Ji et al., 2022; Peng et al., 2023). However, current MLLMs remain inadequate in visual problems that require higher-level

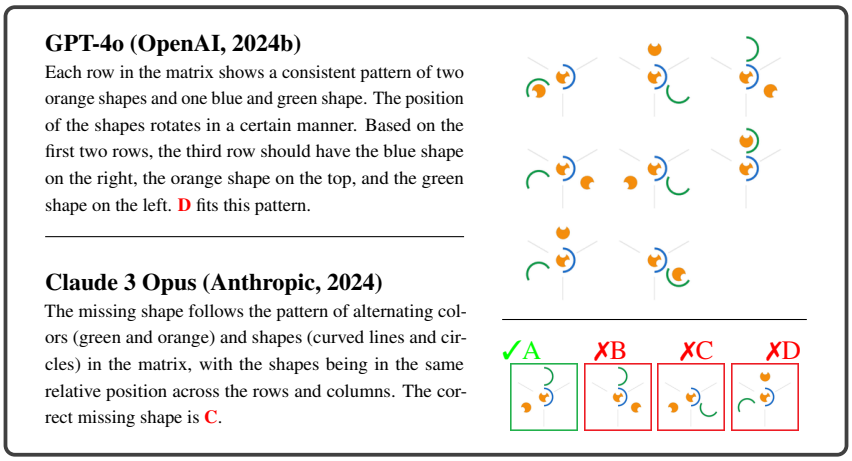


Figure 1: The example of the subpar performance of current state-of-the-art MLLMs (GPT-4o, Claude 3 Opus) on a simple matrix reasoning task used in MaRs-VQA (similar to cases in RPM and WISC). Both models can recognize the basic shapes in the provided patterns but fail to reason the next pattern.

inductive reasoning (Yang et al., 2023). An example is their poor performance on the RAVEN IQ-test (Huang et al., 2024; Fu et al., 2024), which heavily relies on abstract reasoning skills. The RAVEN IQ-test also has some limitations, including a small dataset of only 50 samples (Huang et al., 2024), which may introduce randomness and fail to comprehensively and robustly evaluate MLLMs. Besides, it doesn't include a comparative study with human performance.

To address the ill-posed matrix reasoning assessment and the deficiencies of existing cognitive testing benchmarks, we propose a new VQA dataset – MaRs-VQA, which is the largest psychologist-designed dataset for matrix reasoning assessment including 1,440 examples in total. The sample diversity of MaRs-VQA also surpasses other datasets before. It contains over 50 types of shape, 16 types of colour and over 500 graphic combinations. We also introduce VCog-Bench, the first zero-shot matrix reasoning benchmark to evaluate MLLMs' visual cognition. In VCog-Bench, We conduct thorough evaluation and comparison across 16 existing MLLMs (including their variants) and human performance under a zero-shot inference setting (no prior knowledge) on MaRs-VQA and other abstract reasoning datasets containing human studies. In our experiments, we observe that MLLMs with more parameters generally perform better on our benchmark, adhering to established scaling laws. However, even the largest open-source MLLMs and GPT-4o fall short of surpassing human performance in matrix reasoning tasks. Furthermore, many MLLMs have a mismatch in performance between matrix reasoning tasks and other general VQA problems, which provides some insights into the drawbacks of existing models. In conclusion, our contributions are summarized as follows:

- We introduce a new matrix reasoning VQA dataset – MaRs-VQA, containing 1,440 image instances designed by psychologists, which is the largest dataset for matrix reasoning zero-shot evaluation.
- We propose VCog-Bench, the most comprehensive visual cognition benchmark to date, which evaluates the matrix reasoning performance of 16 existing MLLMs rigorously following the zero-shot setting.
- Our thorough experiments qualitatively reveal the visual cognition gap between MLLMs and humans in matrix reasoning problems. We also show additional insights of deficiencies in MLLMs, which can inspire more future investigations.

2 Related Works

Cognitive Test of Large Language Models (LLMs) The rise of LLMs has aroused interest in exploring human-like AI in psychology and cognition (Ullman, 2023). Recent works tested LLMs' cognitive abilities in causal reasoning (Binz & Schulz, 2023), abstract reasoning (Xu et al., 2023b;

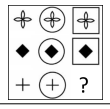
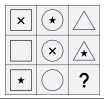
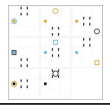
Dataset	Source	Sample	Instance	RGB image	Human Study	Psychological Validity
kosmos-iq50 (NeurIPS-23) (Huang et al., 2024)	RAVEN-IQ Test (Team, b;a; Haven; IQPro.org)		50	✗	✗	✓
Visual Reasoning Benchmark (COLM-24) (Zhang et al., 2024c)	Mensa Test, RAVEN IntelligenceTest (MENSA; Zhang et al., 2019; Labs)		241	✗	✗	✗
MaRs-VQA (ours)	MaRs-IB (Chierchia et al., 2019)		1,440	✓	✓	✓

Table 1: Comparison of recently released zero-shot matrix reasoning datasets to evaluate MLLMs.

Moskvichev et al., 2023; Jiang et al., 2024b; Ahrabian et al., 2024), analogical reasoning (Webb et al., 2023), systematic reasoning (Hagendorff et al., 2023), and theory of mind (Strachan et al., 2024). Their observation showed that LLMs like GPT-4 (Achiam et al., 2023) have been proven successful in most cognitive tests related to language-based reasoning. Despite this success, only limited research has been conducted on the areas of MLLMs and visual cognition. Visual cognition involves the process by which the human visual system interprets and makes inferences about a visual scene using partial information. *Buschoff et al.* observed that while LLMs demonstrate a basic understanding of physical laws and causal relationships, they lack deeper insights into intuitive human preferences and reasoning. Almost all existing visual cognition benchmarks focus on testing MLLMs' cognitive abilities in simple tasks (Lerer et al., 2016; Zhou et al., 2023; Jassim et al., 2023), and ignore testing complex abstract reasoning and logical reasoning ability related to fluid intelligence. Therefore, new and challenging benchmarks based on the theory of visual cognition are needed to assess and improve AI systems' capabilities for human-like visual understanding.

Matrix Reasoning Matrix reasoning is often used to determine human intelligence related to visual cognition and working memory (Salthouse, 1993; Jaeggi et al., 2010; Fleuret et al., 2011) that is widely used by RPM (Raven, 2003; Soulières et al., 2009), WISC (Wechsler & Kodama, 1949; Kaufman et al., 2015) to evaluate human's ability to detect the underlying conceptual relationship among visual objects and use reasoning to find visual cues. Early research indicated that deep learning models can be trained with large-scale matrix reasoning datasets to solve simple matrix reasoning (Stabinger et al., 2021; Małkiński & Mańdziuk, 2022; 2023; Xu et al., 2023a; Małkiński & Mańdziuk, 2024) and compositional visual relation tasks (Fleuret et al., 2011; Zerroug et al., 2022; Ommer & Buhmann, 2007; Liu et al., 2021), achieving human-level accuracy. Several datasets and benchmarks are also proposed, such as PGM (Barrett et al., 2018), RAVEN (Zhang et al., 2019), RAVEN-I (Hu et al., 2021), RAVEN-FAIR (Benny et al., 2021), CVR (Zerroug et al., 2022). However, these works have a key limitation. They ignore that humans can solve these problems by zero-shot reasoning without explicitly learning from large-scale data. After the blooming of LLMs, researchers are keen on testing whether LLMs reached the same abstract reasoning capabilities as humans. *Webb et al.* (Webb et al., 2023) encode matrix reasoning into a symbolic problem based on human's prior and validate LLM can understand this task. Recently, there are also some useful zero-shot visual reasoning inference datasets containing matrix reasoning samples have been proposed in the AI/ML community, such as RAVEN-IQ (Huang et al., 2024) containing 50 instances, Visual Reasoning Benchmark (Zhang et al., 2024c) containing 241 instances in total, but all of them are limited by lacking rigorous human experiments as reference and conducting experiments on relatively small datasets without psychometrical validation.

Vision-Language Models Researchers have been actively investigating the utility of Vision-Language Models (VLMs) for addressing vision reasoning tasks (Zellers et al., 2019; Bordes et al., 2024). These latest VLMs are constructed using a combination of the CLIP vision encoder, pre-trained LLMs, and a connected adapter to align visual features with language space (Zhang et al., 2024b; Shao et al., 2024; Gupta & Kembhavi, 2023; Fu et al., 2024). Notably, methodologies

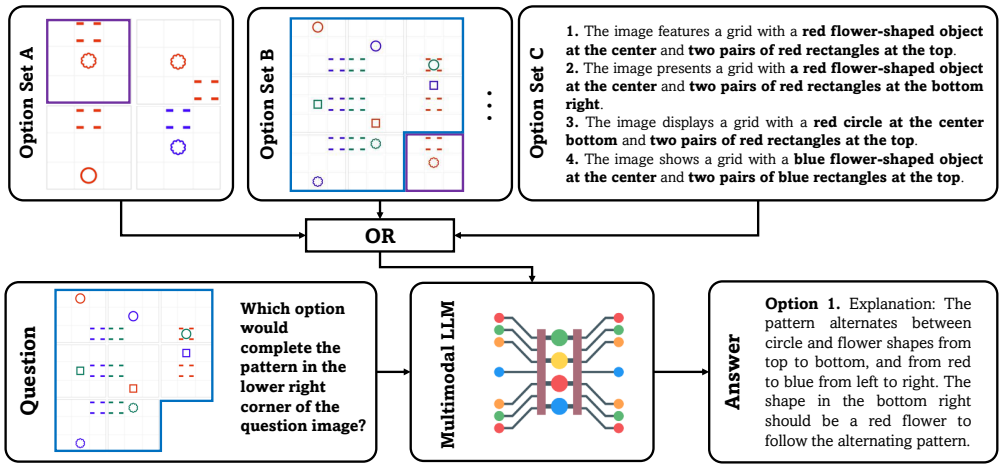


Figure 2: An example of question and option sets in MaRs-VQA to evaluate Multimodal LLMs. The input set contains an image with a corresponding question and three sets of four-option images/contexts. Option Set A includes single-object images that can be filled into the blank region. Option Set B includes full 3x3 images containing all objects. Option C includes language descriptions for each option.

such as MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2024), LLaVA (Liu et al., 2024b), CogVLM (Wang et al., 2023) underscore the significance of employing high-quality visual instruction tuning data. Additionally, tool learning methods have also explored the potential of integrating code generation pipelines with visual inference (Surís et al., 2023). Nevertheless, current VLMs encounter challenges in adapting to high-resolution and visually complex images. These problems stem from the absence of a robust visual search mechanism (Wu & Xie, 2023), few-shot reasoning (Guo et al., 2023), compositional understanding (Yuksekgonul et al., 2022) and the constrained visual grounding capabilities inherent in CLIP (Tong et al., 2024).

3 MARS-VQA DATASET

The MaRs-VQA dataset is designed to evaluate the zero-shot abstract reasoning capabilities of MLLMs through various matrix reasoning VQA tasks. All sample images in MaRs-VQA are sourced from the Matrix Reasoning Item Bank (MaRs-IB) (Chierchia et al., 2019), which is created by psychologists including 18 sets of abstract reasoning questionnaires (80 instances in each set) for non-verbal abstract reasoning assessment of adolescents and adults. Each item presents an incomplete 3×3 matrix of abstract shapes, requiring participants to identify relationships among the shapes. Compared to RAVEN and other matrix reasoning created by computer scientists, the matrix reasoning samples in MaRs-VQA are psychometrically validated and widely used in neurodevelopmental and neuropsychological research (Keating et al., 2022; Zorowitz et al., 2024; Nussenbaum et al., 2020; Moses-Payne et al., 2022).

In Figure 2, we demonstrate how to transform the matrix reasoning problem into a VQA task for LLMs and VLMs using a sample from the MaRs-VQA dataset. We define three different option sets – two image-based sets (A and B) and one language-based set (C). In Option Set A, we provide four candidates to the missing patch in the question. In Option Set B, the options are created by filling the four patches in Set A into the 3×3 question image. Note that Option Set B is used for visualization purposes only and is not included in our experiment. We further diversify the modalities of our dataset to support the evaluation of different kinds of models. Specifically, we use GPT-4o and human annotators to generate language-based descriptions for each option, forming Option Set C. In the data generation process, we first manually design 10 VQA examples, which serve as the initial human annotations in our data collection. These examples are then used as few-shot samples to query GPT-4o through in-context learning. The context generation system prompt guides GPT-4o to

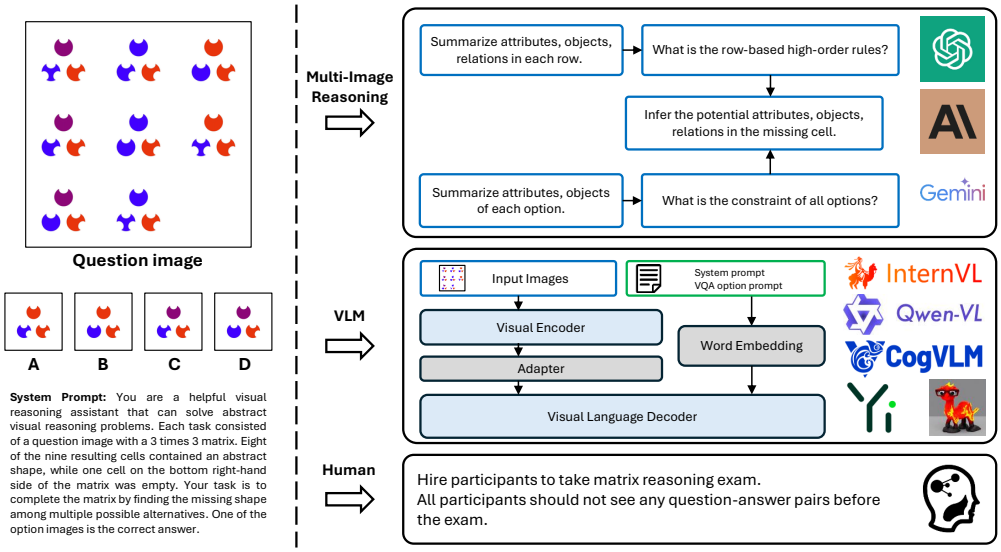


Figure 3: An overview of the VCog-Bench. The left part is the model input, including a question image, multiple option images and a system prompt describing the task. The right part shows the step-by-step CoT for multi-image reasoning and VLM solution for matrix reasoning problems.

compare all four option images and generate distinct descriptions for each one. After generating all samples, human annotators in the author team review each option and revise the incorrect description.

4 Visual Cognition Benchmark (VCog-Bench)

Different from the original setting in other matrix reasoning dataset like RAVEN (Zhang et al., 2019), our goal of MLLM agent in VCog-Bench is to complete the 3×3 matrix by finding the missing cell from multiple options by zero-shot learning. To this end, MLLM agents have to deduce relationships across the other cells of the matrix and infer the missing cell accordingly. Based on the current progress of Multimodal LLMs, we propose two potential solutions.

4.1 Multi-Image Reasoning via Chain-of-Thought (CoT)

Recent research in the NLP community has revealed the effectiveness of CoT in improving the reasoning capability of LLMs for complex problems (Wei et al., 2022; Kojima et al., 2022). In this paper, we propose the object-centric CoT prompting strategy, which combines the ideas of CoT (Zhang et al., 2023; Zhou et al., 2024; Zhang et al., 2024a), object-centric relational abstraction (Webb et al., 2024a;b; Mondal et al., 2024; Xu et al., 2023b) and object-centric representation learning (Seitzer et al., 2022; Dittadi et al., 2022; Jiang et al., 2024a), to enhance the MLLM's zero-shot learning performance in solving matrix reasoning problems.

Following previous works (Carpenter et al., 1990; Barrett et al., 2018; Chierchia et al., 2019; Zhang et al., 2019), we formulate the structure K of matrix reasoning as a combination of four components, $K = \{[r, a, o, s] | r \in \mathcal{R}, a \in \mathcal{A}, o \in \mathcal{O}, s \in \mathcal{S}\}$. $\mathcal{R}$ is a set of rules of how the pattern changes along each row and column (*e.g.*, rotating by a fixed angle and shifting by a fixed distance); $\mathcal{A}$ is a set of attributes in each pattern (*e.g.*, color, shape, and size); $\mathcal{O}$ is how to integrate objects in each cell (*e.g.*, spatial location and overlap); $\mathcal{S}$ denotes a set of constraints for designing answer options (*e.g.*, options should have minimum difference), which avoids that participants solving the matrix reasoning problems in unintended ways.

Based on structure K, we use three steps to guide MLLM to use human-level thought to understand matrix reasoning tasks. The first step is to guide the Multimodal LLM to summarize the visual feature (e.g. shape) of each row in the 3×3 question image. Then, based on these row-based visual features, the model will then conclude the high-order rule/pattern $\mathcal{R}$. The second step is to extract

the basic attributes $\mathcal{A}$ and inner relations $\mathcal{O}$ to integrate objects in each option image. The third step is to infer the answer based on exclusion with potential answer designed constraints $\mathcal{S}$. The Multi-Image Reasoning section of Figure 3 shows a schematic depiction of how to leverage CoT in matrix reasoning tasks.

4.2 Vision-Language Models (VLMs)

In addition to MLLMs, we also evaluate the performance of VLMs for a thorough comparison. In VLMs, we only use question image as visual input and transform all option images into language descriptions (*i.e.*, Option Set C), which matches the input representations required by VLMs (Xu et al., 2023b; Camposampiero et al., 2023). The VLM section in Figure 3 illustrates this pipeline.

The test set contains n VQA samples, denoted as $\{(\mathbf{q}_i, \mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$. $\mathbf{q}_i$ represents the question image showing the 3×3 matrix reasoning task (MaRs-VQA). $\mathbf{x}_i = [x_i^1, ..., x_i^k]$ represents the context description in the option set, where k is the number of options. $\mathbf{y}_i$ is the answer of the matrix reasoning question. The zero-shot inference pipeline of VLM can be formulated as:

$$\hat{\mathbf{y}}_i = F_\theta(\mathbf{q}_i, \mathbf{x}_i, \mathbf{x}_{sys}). \tag{1}$$

$\mathbf{x}_{sys}$ is the system prompt, including independent information about the matrix reasoning problem setting, structure K for each dataset and requirements for the output format. $\hat{\mathbf{y}}_i$ is the prediction result .F_θ is an autoregressive decoder in the LLM for answer generation. It is defined as:

$$P(\hat{\mathbf{y}}_i|\mathbf{q}_i, \mathbf{x}_i, \mathbf{x}_{sys}) = \prod_{j=1}^{L} P(\hat{\mathbf{y}}_{i,j}|f(\mathbf{q}_i), \mathbf{x}_i, \mathbf{x}_{sys}, \hat{\mathbf{y}}_{i,<j}; \theta), \tag{2}$$

where f is the visual encoder and adapter layer, L is the sequence length of answers and $\hat{\mathbf{y}}_{i,<j}$ is all answer tokens before $\hat{\mathbf{y}}_{i,j}$.

In VLMs, the input question image is first processed by the visual encoder such as CLIP (Radford et al., 2021). Then, additional adapter layers are used to map visual features into language feature space. These features, along with the context-based option descriptions, are sent to the LLM decoder. The LLM decoder then integrates the information from both the input question image and the option descriptions to address the VQA task. VLMs leverage the strengths of both visual encoders and language models, allowing for a more comprehensive analysis of the matrix reasoning problems. It provides a structured way to break down the problem, potentially improving interpretability compared to end-to-end close-source models.

5 Experiments

5.1 Experimental Settings

Datasets In addition to MaRs-VQA, we selected two well-known open-source datasets for matrix reasoning and abstract visual reasoning to conduct experiments in VCog-Bench. The first dataset is RAVEN (Zhang et al., 2019), designed to probe abstract reasoning in a format similar to the Raven's Progressive Matrices IQ test, with each question providing eight options. The second dataset is Compositional Visual Reasoning (CVR) (Zerroug et al., 2022), which evaluates deep learning models using 103 unique configurations generated by predefined rules. Each sample in CVR is an outlier detection problem, with four options provided per question. However, both RAVEN and CVR share a significant limitation: all samples are algorithmically generated using fixed rules, which limits their diversity and lacks psychological validity.

Baselines for Multi-image Reasoning We selected the Claude 3 family (Haiku, Sonnet, Opus) (Anthropic, 2024), GPT-4V (OpenAI, 2023), GPT-4o (OpenAI, 2024b) as the primary multi-image CoT baselines as they support multiple images input, so they are tested with a more difficult setting in Table 2, *i.e.*, the input is a question and multiple option images in Option Set A of Figure 2.

Method	Learning	Accuracy (%) ↑		
		MaRs-VQA (4-options)	RAVEN (8-options)	CVR (4-options)
Claude 3 Sonnet (Anthropic, 2024)	zero-shot	22.92	10.71	27.83
	chain-of-thought	23.22	13.39	28.48
Claude 3 Opus (Anthropic, 2024)	zero-shot	20.85	11.61	26.86
	chain-of-thought	24.13	11.95	27.18
Claude 3.5 Sonnet (Anthropic, 2024)	zero-shot	23.18	14.08	25.97
	chain-of-thought	24.28	15.36	27.88
GPT-4V (OpenAI, 2023)	zero-shot	27.71	13.84	36.25
	chain-of-thought	33.13	15.63	40.62
GPT-4o (OpenAI, 2024b)	zero-shot	30.21	19.20	42.50
	chain-of-thought	33.96	25.89	44.01
Human	-	**69.15**	**84.41**	**78.70**

Table 2: Experiments on multi-image reasoning. zero-shot means only provide the model system prompt about the matrix reasoning task definition. Chain-of-thought denotes the implementation in section 4.1. The results are averaged over three runs with three different random seeds.

Baselines for VLMs For the VLMs, we select state-of-the-arts open-source and closed-source models such as InstructBLIP (Dai et al., 2024), MiniGPT-v2 (Zhu et al., 2023), LLaVA-v1.6 (LLaVA-NeXT) (Liu et al., 2024a), CogVLMv2 (Wang et al., 2023), Yi-VL (Young et al., 2024), Qwen-VL (Bai et al., 2023), InternVL (Chen et al., 2024), Gemini Pro 1.5 (Reid et al., 2024), Claude 3 family (Haiku, Sonnet, Opus) (Anthropic, 2024), GPT-4V (OpenAI, 2023), GPT-4o (OpenAI, 2024b) as the primary VLM baselines. The input is a question image and language-based options.

Human Baseline The human study results in Table 2 and 3 are reported from previous experiment results. The human subjects of RAVEN (Zhang et al., 2019) consists of college students from a subject pool maintained by the Department of Psychology. Only "easily perceptible" examples were used in the investigation. CVR (Zerroug et al., 2022) hired 21 participants and each participant completed 6 different tasks with 20 problem samples for each task. The human study results of MaRs-IB (Chierchia et al., 2019) (data source of MaRs-VQA) are more rigorous. They are from 4 age groups ($N = 659$, aged 11–33 years). The accuracy for younger adolescents, mid-adolescents, older adolescents, and adults solving matrix reasoning in MaRs-IB are 61%, 68%, 73%, 81%. We use the average result of all groups in Table 2 and 3.

Implementation For closed-source baseline models, we establish basic prompts to introduce the matrix reasoning problem setting, which serve as the system prompt for zero-shot inference. For object-centric CoT reasoning, we create specific prompts to guide the model's thought process through multiple stages, enabling step-by-step reasoning. For open-source baseline models, we use the same system prompt settings across all models. Testing is conducted using two NVIDIA RTX 4090 GPUs for 7B-sized VLMs and eight NVIDIA A100 80GB GPUs for VLMs larger than 7B. All experiments are run with three different random seeds, and the results are averaged. We evaluate the results based on the accuracy of single-option matrix reasoning problems (Acc $=$ Correct/Total), consistent with other VQA benchmarks (Lu et al., 2022; Liu et al., 2023).

5.2 Experimental Results

In this subsection, we present the experimental results of the baselines in the VCog-Bench. The results demonstrate that while parts of baseline models can understand some basic forms of the matrix reasoning task, they struggle with complex tasks requiring both visual working memory and multi-image reasoning capability.

We divided our experiments into two parts. The first part involves end-to-end multi-image reasoning. For this experiment, we used multiple images as the input, including a question image and several option images (refer to Option Set A in Figure 2), and guided the MLLMs to decompose the problem into predefined structures before generating answers based on all available information. We tested the Claude 3 family, GPT-4V, and GPT-4o for this task, as these models support multi-image reasoning. Table 2 shows that even the state-of-the-art closed-source MLLMs perform worse than humans in all matrix reasoning tasks. While object-centric CoT can help larger models achieve better performance,

Method	Training Data	Model Scale	LLM Backbone	Accuracy (%) ↑	
				MaRs-VQA (4 Options)	RAVEN (8 Options)
InstructBLIP (Dai et al., 2024)	129M	7B	Vicuna-7B (Chiang et al., 2023)	10.63	12.05
LLaVA-v1.6 (Liu et al., 2024b)	1.3M	7B	Mistral-7B (Jiang et al., 2023)	16.88	14.29
MiniGPT-v2 (Zhu et al., 2023)	-	8B	Llama-2-7B (Touvron et al., 2023)	26.45	13.39
Qwen-VL (Bai et al., 2023)	1.4B	10B	Qwen-7B (Bai et al., 2023)	29.58	16.07
InstructBLIP (Dai et al., 2024)	129M	13B	Vicuna-13B (Chiang et al., 2023)	10.42	14.46
CogVLMv2 (Wang et al., 2023)	1.5B	19B	Llama-3-8B (Meta, 2024a)	26.46	12.05
InternVL 1.5 (Chen et al., 2024)	6.0B	26B	InternLM2-Chat-20B (Cai et al., 2024)	22.09	14.73
Yi-VL (Young et al., 2024)	100M	34B	Yi-34B-Chat (Young et al., 2024)	25.21	19.64
LLaVA-v1.6 (Liu et al., 2024b)	1.3M	35B	Hermes-Yi-34B (Young et al., 2024)	34.38	33.93
InternVL 1.2+ (Chen et al., 2024)	6.0B	40B	Hermes-Yi-34B (Young et al., 2024)	32.71	33.04
Qwen2-VL (Wang et al., 2024)	-	72B	Qwen2-72B (Yang et al., 2024)	34.22	36.15
InternVL 2 (Chen et al., 2024)	-	76B	Hermes-2-Theta-Llama-3-70B (Teknium et al.)	34.63	38.01
Llama 3.2 (Meta, 2024b)	6.0B	90B	-	34.81	35.26
Claude 3.5 Sonnet (Anthropic, 2024)	*unknown*	*unknown*	*unknown*	34.82	35.36
GPT-4o (OpenAI, 2024b)	*unknown*	*unknown*	*unknown*	37.38	38.84
Gemini Pro 1.5 (Reid et al., 2024)	*unknown*	*unknown*	*unknown*	34.79	42.86
Human	-	-	-	**69.15**	**84.41**

Table 3: Experiments on using a question image and language descriptions for options as inputs to compare different VLMs. The results are averaged over three random seeds.

Method	Multi-Image	Accuracy (%) ↑				
		Level 1 (90)	Level 2 (96)	Level 3 (84)	Level 4 (72)	Level >4 (138)
Claude 3 Opus (Anthropic, 2024)	✓	19.15	28.57	13.34	13.16	24.66
GPT-4o (OpenAI, 2024b)	✓	57.78	27.08	27.38	19.43	21.74
Claude 3 Opus (Anthropic, 2024)	✗	24.44	25.00	40.48	38.89	39.13
Gemini Pro 1.5 (Reid et al., 2024)	✗	51.10	30.21	26.19	29.17	35.51
GPT-4o (OpenAI, 2024b)	✗	58.89	45.83	32.14	26.39	26.09

Table 4: Compare closed-source MLLMs 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.

it does not benefit smaller models such as Claude 3 Sonnet. Compared to the results in MaRs-VQA and RAVEN, GPT-4o achieves much better zero-shot and object-centric CoT inference results in the CVR dataset, almost matching the performance (ResNet-50: 57.9%, ViT-small: 32.7%, WReN: 42.4%) of fine-tuned models with 1,000 training samples in CVR's paper (Zerroug et al., 2022).

In the second part of our experiment, we investigated the use of VLMs (question image + language options) to solve matrix reasoning problems in MaRs-VQA and RAVEN. The CVR dataset was excluded because the shapes it contains are too complex to describe accurately. As shown in Table 3, large-scale VLMs, such as Qwen2-72B and InternVL-2-76B, achieved comparable results to GPT-4o in MaRs-VQA and RAVEN. Notably, Gemini Pro 1.5 outperformed GPT-4o on the RAVEN dataset.

We identified three major issues after reviewing the reasoning outputs of current MLLMs in Table 2 and 3: (1) Limited Use of Visual Information: MLLMs cannot directly use visual features for reasoning, making them insensitive to non-verbal spatial features during CoT reasoning. This limitation is particularly evident when handling images that require describing the positional relations of objects. For example, it is difficult for MLLMs to distinguish each option in Figure 1 using language alone. (2) Restricted Visual Working Memory: The visual working memory of MLLMs is limited, causing visual feature information to be easily lost during the text generation reasoning process. (3) Integration Challenges: Even if MLLMs possess strong task-specific skills like recognition, segmentation, and object detection, they struggle to integrate these skills into high-level visual reasoning tasks.

5.3 Ablation Study and Qualitative Analysis

In this subsection, we use case studies from the MaRs-VQA dataset to illustrate how MLLMs fail in some matrix reasoning tasks and provide insights on how to improve MLLMs and VLMs for this task. We also compare different MLLMs with difficulty levels in the MaRs-VQA dataset (see Table 4).

First, we present an example to explain why the Claude 3 family performs worse than GPT-4o and even worse than random guessing in all of our experiments. Figure 4 top left is one of the most simple cases in MaRs-VQA's level 1 difficulty, Claude 3 Opus incorrectly identifies the shape as the main target of this matrix, while the actual target is the size. In contrast, GPT-4o correctly discerns

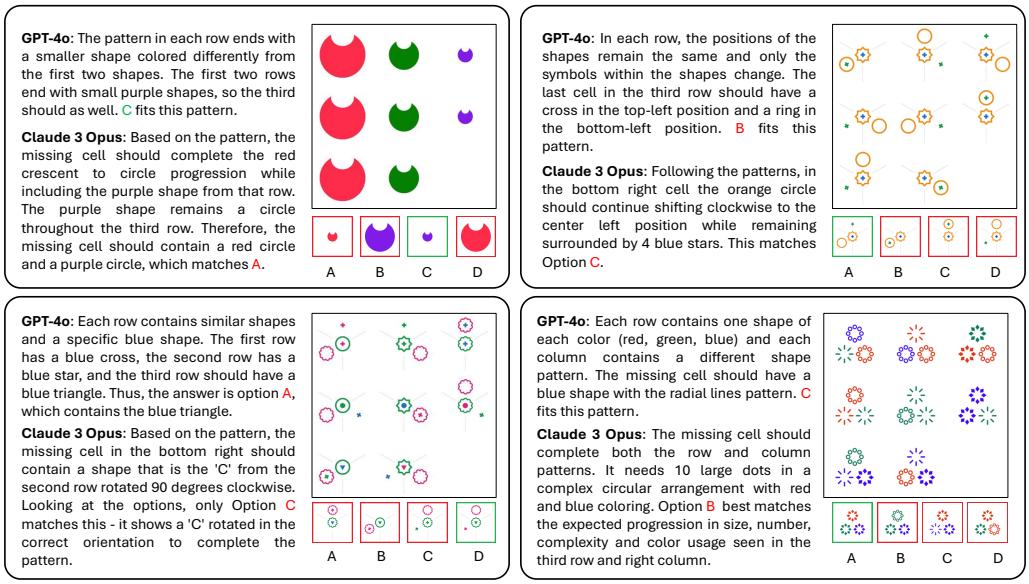


Figure 4: Different matrix reasoning problem (difficulty levels) from MaRs-VQA and MLLM's reply. We use green to represent correct answer and red to represent wrong answer of each question. The top left is a sample with difficulty level 1. The others are samples with difficulty level $\geq$4.

the relationship between rows, noting: "The pattern in each row ends with a smaller shape colored differently from the first two shapes." This example highlights a critical shortcoming in Claude 3 Opus's reasoning ability: limited Use of Visual Information, demonstrating its struggle to accurately interpret the key attributes in matrix reasoning tasks. GPT-4o, on the other hand, showcases a superior understanding of the relationships and patterns within the data, leading to more accurate responses.

However, both GPT-4o and Claude 3 Opus are failed with too difficult cases ($\geq$ level 4), Figure 4 bottom left and shows an example, it is the level 6 difficulty containing shape, positional relation, shape with different objects. The MLLMs are easy to hallucinate on these tasks. Note: the reasoning here is a short summary of the chain-of-thought, not the full version. The right part of Figure 4 illustrates instances where GPT-4o failed due to poor utilization of visual features and a lack of visual working memory. Additionally, we observed that GPT-4o is not sensitive to the positional relationships in the question images.

These failures highlight significant limitations in MLLM's visual processing capabilities. The model's inability to effectively leverage visual features and its lack of visual working memory result in incorrect interpretations. Furthermore, its insensitivity to positional relationships underscores a critical area for improvement in understanding and analyzing spatial information in visual reasoning.

5.4 VISUALIZATION

We also analyze the relationship between matrix reasoning accuracy and model scale in Figure 5. The figure illustrates the significant gap between MLLM's matrix reasoning performance and that of humans. This gap is substantial and suggests that simply increasing model size according to scaling laws will not be sufficient to bridge it.

Although LLMs have achieved remarkable success in language understanding and generation, a significant portion of their parameters is dedicated to encoding linguistic patterns and memorizing factual information. This focus offers limited benefits for tasks requiring visual cognition. This disparity between Multimodal LLMs and human indicates that merely increasing model size is insufficient to achieve human-level zero-shot inference in these domains. Drawing inspiration from recent strategies proposed by OpenAI o1 (OpenAI, 2024a), a potential solution is to shift computational resources from extensive LLM training to enhancing the inference process. By conceptualizing LLMs as text-based simulators and iteratively rolling out problem-solving strategies, models can potentially converge on effective solutions in a manner analogous to human reasoning.

6 DISCUSSION

Social Impacts In the present work, we emphasize that zero-shot matrix reasoning is a key item to validate human-level intelligence, though it is still unclear how matrix reasoning ability is acquired early in human neurodevelopment. Children's visual reasoners (without any additional training) can provide sensible answers to matrix reasoning questions as early as age four. The long-term goal of our work is twofold. The first one is to explore the problem of how close AIs or MLLMs are to human-like cognitive abilities, which is raised by *François Chollet* in 2019 Chollet (2019). The second one is to develop an MLLM-powered AI agent that can simulate human-level zero-shot matrix reasoning capability. The agent will eventually guide vision generation models to generate new matrix reasoning samples and tasks and design new neurodevelopmental assessment tools. This will help psychologists and pediatricians explore and deconstruct how children activate such abilities in the early stage of neurodevelopment.

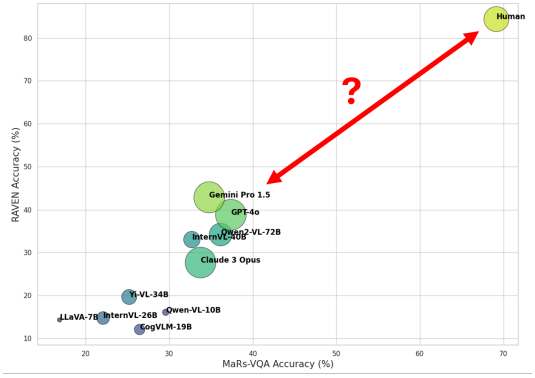


Figure 5: There is still a big gap between human's matrix reasoning capability and MLLM's. Bubble size corresponds to the model size. As we don't know the exact size of closed-source MLLMs, we set all of them to the largest value by default. The model size of human refers to the number of neurons (86B) in human's brain (Voytek, 2013).

Limitations An open-ended question is whether MLLMs need to achieve or surpass human-level zero-shot inference capability in matrix reasoning tasks. Addressing this issue requires new theories from cognitive science and psychology to accurately evaluate and compare human and MLLM intelligence. Unlike MLLMs, which rely on training data and domain-specific skills, human cognition develops gradually and evolves with age. Humans can also learn how to solve the problem progressively from previous seen matrix reasoning tasks while they are taking the test. Therefore, AI researchers, psychologists, and cognitive scientists must collaborate to rethink how to benchmark MLLM intelligence with human intelligence.

Ethics 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), and VCog-Bench builds on MaRs-VQA, RAVEN (GPL-3.0 License), CVR (Apache License 2.0). All code and data are available on GitHub. No conflicts of interest exist among the study's contributors. More discussion on the ethical aspects of VCog-Bench is included in the Appendix. The annotation process is IRB approved by a clinical institute.

7 CONCLUSION

We introduce VCog-Bench, a publicly available zero-shot matrix reasoning benchmark designed to evaluate the visual cognition capability and intelligence of Multimodal Large Language Models (MLLMs). This benchmark integrates two well-known datasets RAVEN and CVR from the AI community and includes our newly proposed MaRs-VQA dataset. We also introduce several important concepts to redefine zero-shot matrix reasoning task evaluation, focusing on multi-image reasoning with object-centric Chain-of-Thought (CoT) system prompts. Our findings show that current state-of-the-art MLLMs and Vision-Language Models (VLMs), such as GPT-4o and LLaVA-1.6, InternVL demonstrate some basic understanding of matrix reasoning tasks. However, these models still face big challenges with complex situations and perform much worse than human. This highlights the need for further exploration and development in this area. By providing a robust benchmark, we aim to encourage further innovation and progress in the field of improving the visual cognition of MLLMs.

REFERENCES

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

Kian Ahrabian, Zhivar Sourati, Kexuan Sun, Jiarui Zhang, Yifan Jiang, Fred Morstatter, and Jay Pujara. The curious case of nonverbal abstract reasoning with multi-modal large language models. *arXiv preprint arXiv:2401.12117*, 2024.

Anthropic. Introducing the next generation of claude. `https://www.anthropic.com/news/claude-3-family`, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pp. 511–520. PMLR, 2018.

Yaniv Benny, Niv Pekar, and Lior Wolf. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12557–12565, 2021.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Giacomo Camposampiero, Loïc Houmard, Benjamin Estermann, Joël Mathys, and Roger Wattenhofer. Abstract visual reasoning enabled by language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2642–2646, 2023.

Patricia A Carpenter, Marcel A Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97(3):404, 1990.

Raymond Bernard Cattell and Alberta KS Cattell. *Measuring intelligence with the culture fair tests*. Institute for Personality and Ability Testing, 1960.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Gabriele Chierchia, Delia Fuhrmann, Lisa J Knoll, Blanca Piera Pi-Sunyer, Ashok L Sakhardande, and Sarah-Jayne Blakemore. The matrix reasoning item bank (mars-ib): novel, open-access abstract reasoning items for adolescents and adults. *Royal Society open science*, 6(10):190232, 2019.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

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

Andrea Dittadi, Samuele S Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. Generalization and robustness implications in object-centric learning. In *International Conference on Machine Learning*, pp. 5221–5285. PMLR, 2022.

François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

Dedre Gentner. Children's performance on a spatial analogies task. *Child development*, pp. 1034–1039, 1977.

Qing Guo, Prashan Wanigasekara, Jian Zheng, Jacob Zhiyuan Fang, Xinwei Deng, and Chenyang Tao. How do large multimodal models really fare in classical vision few-shot challenges? a deep dive. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838, 2023.

IQ Haven. Matrix-g iq test. `https://iqhaven.com/matrix-g`.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1567–1574, 2021.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024.

IQPro.org. Culturally fair nonverbal iq test built on science. `https://iqpro.org/`.

Susanne M Jaeggi, Barbara Studer-Luethi, Martin Buschkuehl, Yi-Fen Su, John Jonides, and Walter J Perrig. The relationship between n-back performance and matrix reasoning—implications for training and transfer. *Intelligence*, 38(6):625–635, 2010.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*, 2023.

Arthur R Jensen. The factor. *Westport, CT: Prager*, 1998.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. *arXiv preprint arXiv:2211.16492*, 2022.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *Advances in Neural Information Processing Systems*, 36, 2024a.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *arXiv preprint arXiv:2404.13591*, 2024b.

Alan S Kaufman, Susan Engi Raiford, and Diane L Coalson. *Intelligent testing with the WISC-V*. John Wiley & Sons, 2015.

Connor T Keating, Dagmar S Fraser, Sophie Sowden, and Jennifer L Cook. Differences between autistic and non-autistic adults in the recognition of anger from facial motion remain after controlling for alexithymia. *Journal of autism and developmental disorders*, 52(4):1855–1871, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

IQ Test Labs. Iq test - 20 questions. `https://www.intelligencetest.com/questions/`.

Paulo Guirro Laurence and Elizeu Coutinho Macedo. Cognitive strategies in matrix-reasoning tasks: State of the art. *Psychonomic Bulletin & Review*, 30(1):147–159, 2023.

Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International conference on machine learning*, pp. 430–438. PMLR, 2016.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next`, 2024a.

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

Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, and Antonio Torralba. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34:23166–23178, 2021.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Mikołaj Małkiński and Jacek Mańdziuk. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *arXiv preprint arXiv:2201.12382*, 2022.

Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736, 2023.

Mikołaj Małkiński and Jacek Mańdziuk. One self-configurable model to solve many abstract visual reasoning problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14297–14305, 2024.

MENSA. Mensa iq challenge. `https://www.mensa.org/mensa-iq-challenge/`.

AI Meta. Introducing meta llama 3: The most capable openly available llm to date, 2024. *URL https://ai. meta. com/blog/meta-llama-3/. Accessed on April*, 26, 2024a.

AI Meta. Introducing llama 3.2. *URL https://github.com/meta-llama/llama-models/tree/main/models/llama3_2 Accessed on Sep*, 2024b.

Shanka Subhra Mondal, Jonathan D Cohen, and Taylor W Webb. Slot abstractors: Toward scalable abstract visual reasoning. *arXiv preprint arXiv:2403.03458*, 2024.

ME Moses-Payne, G Chierchia, and S-J Blakemore. Age-related changes in the impact of valence on self-referential processing in female adolescents and young adults. *Cognitive Development*, 61: 101128, 2022.

Arsenii Kirillovich Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *Transactions on Machine Learning Research*, 2023.

Kate Nussenbaum, Maximilian Scheuplein, Camille V Phaneuf, Michael D Evans, and Catherine A Hartley. Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology*, 6(1), 2020.

Bjorn Ommer and Joachim M Buhmann. Learning the compositional nature of visual objects. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

OpenAI. Gpt-4v(ision) system card. `https://openai.com/research/gpt-4v-system-card`, 2023.

OpenAI. Learning to reason with llms. `https://openai.com/index/learning-to-reason-with-llms/`, 2024a.

OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o`, 2024b.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Jean Raven. Raven progressive matrices. In *Handbook of nonverbal assessment*, pp. 223–237. Springer, 2003.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Timothy A Salthouse. Influence of working memory on adult age differences in matrix reasoning. *British Journal of Psychology*, 84(2):171–199, 1993.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

Isabelle Soulières, Michelle Dawson, Fabienne Samson, Elise B Barbeau, Cherif P Sahyoun, Gary E Strangman, Thomas A Zeffiro, and Laurent Mottron. Enhanced visual processing contributes to matrix reasoning in autism. *Human brain mapping*, 30(12):4082–4107, 2009.

Sebastian Stabinger, David Peer, Justus Piater, and Antonio Rodríguez-Sánchez. Evaluating the progress of deep learning for visual relational concepts. *Journal of Vision*, 21(11):8–8, 2021.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pp. 1–11, 2024.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

Testometrika Team. Iq test. https://en.testometrika.com/intellectual/iq-test/, a.

Testometrika Team. Iq test for kids from 7 to 16 year old. https://en.testometrika.com/intellectual/iq-test-for-kids-7-to-16-year-old/, b.

Teknium, Charles Goddard, interstellarninja, theemozilla, karan4d, and huemin_art. Hermes-2-theta-llama-3-70b. https://huggingface.co/NousResearch/Hermes-2-Theta-Llama-3-70B.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Bradley Voytek. Are there really as many neurons in the human brain as stars in the milky way. *Scitable, Nature Education*, 2013.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

Taylor Webb, Shanka Subhra Mondal, and Jonathan D Cohen. Systematic visual reasoning through object-centric relational abstraction. *Advances in Neural Information Processing Systems*, 36, 2024a.

Taylor W Webb, Steven M Frankland, Awni Altabaa, Simon Segert, Kamesh Krishnamurthy, Declan Campbell, Jacob Russin, Tyler Giallanza, Randall O'Reilly, John Lafferty, et al. The relational bottleneck as an inductive bias for efficient abstraction. *Trends in Cognitive Sciences*, 2024b.

Taylor Whittington Webb, Ishan Sinha, and Jonathan Cohen. Emergent symbols through binding in external memory. In *International Conference on Learning Representations*, 2020.

David Wechsler and Habuku Kodama. *Wechsler intelligence scale for children*, volume 1. Psychological corporation New York, 1949.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jingyi Xu, Tushar Vaidya, Yufei Wu, Saket Chandra, Zhangsheng Lai, and Kai Fong Ernest Chong. Abstract visual reasoning: An algebraic approach for solving raven's progressive matrices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6715–6724, 2023a.

Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*, 2023b.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *Advances in neural information processing systems*, 35: 29776–29788, 2022.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5317–5327, 2019.

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

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024b.

Yizhe Zhang, He Bai, Ruixiang Zhang, Jiatao Gu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. How far are we from intelligent visual deductive reasoning? *arXiv preprint arXiv:2403.04732*, 2024c.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Kai Zhao, Chang Xu, and Bailu Si. Learning visual abstract reasoning through dual-stream networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16979–16988, 2024.

Liang Zhou, Kevin A Smith, Joshua B Tenenbaum, and Tobias Gerstenberg. Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*, 152(8):2237, 2023.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Samuel Zorowitz, Gabriele Chierchia, Sarah-Jayne Blakemore, and Nathaniel D Daw. An item response theory analysis of the matrix reasoning item bank (mars-ib). *Behavior research methods*, 56(3):1104–1122, 2024.

Appendices

CONTENTS

1 Introduction **1**

2 Related Works **2**

3 MaRs-VQA Dataset **4**

4 Visual Cognition Benchmark (VCog-Bench) **5**

4.1 Multi-Image Reasoning via Chain-of-Thought (CoT) 5

4.2 Vision-Language Models (VLMs) 6

5 Experiments **6**

5.1 Experimental Settings 6

5.2 Experimental Results 7

5.3 Ablation Study and Qualitative Analysis 8

5.4 Visualization 9

6 Discussion **10**

7 Conclusion **10**

Appendices **17**

A Datasets & Benchmarking Code **18**

B Data Collection and Licenses **18**

C Experimental Settings **19**

C.1 Implementation Details 19

C.2 System Prompts 19

D More Discussion on Limitations and Future Work **20**

E Ethics Discussion **22**

E.1 Negative Societal Impacts 22

E.2 Mitigating Bias and Negative Societal Impacts 22

A DATASETS & BENCHMARKING CODE

We release the data and annotations of MaRs-VQA anonymously:

huggingface.co/datasets/vcog/marsvqa

We also release the code for Multimodal Large Language Models (MLLM) inference in an anonymous github repo:

anonymous.4open.science/r/VCog-Bench-94D2

B DATA COLLECTION AND LICENSES

We showed and compared all datasets in VCog-Bench in Table 6. The data collection of VCog-Bench follows strict procedures. The reason we choose RAVEN, CVR, MaRs-VQA is because all these datasets contain zero-shot / few-shot human investigation results. Based on these results, we can compare the MLLM's performance with human in matrix reasoning tasks.

For RAVEN and CVR, 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 in VCog-Bench is 560.

CVR The original dataset link of CVR is github.com/serre-lab/CVR. It is under Apache License 2.0 (CVR LICENSE). CVR is an accepted paper by NeurIPS 2022 Datasets and Benchmarks track, so all of its data is free to use by public. We follow the same data generation pipeline in CVR to generate 309 samples.

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.

After we download all questionnaires from MaRs-IB, we use two Python scripts to merge all question-option pairs from different questionnaires into the same sample set. Then, we generate Option Set A, Option Set B in Figure 2 by manipulating the size and image position of option images. After that, we annotate the language description of 4 options in 10 samples from the raw data. The language description is used as system prompt to guide GPT-4o to generate all description for all data in MaRs-VQA. Then, human annotators review the annotation and revise them. Finally, we publish all annotations as Option Set A, Option Set B, and Option Set C for MaRs-VQA.

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.

C Experimental Settings

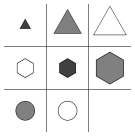
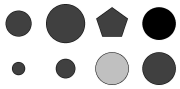
Dataset	Question	Option	Instance	Description
RAVEN (Zhang et al., 2019)			rule-based generation	8 options per instance grayscale image rule-based stimuli **include human study**
CVR (Zerroug et al., 2022)	**Find the outlier among 4 images**		rule-based generation	4 options per instance RGB image rule-based stimuli **include human study**
MaRs-VQA			1,440	4 options per instance RGB image psychologist designed stimuli **include human study**

Table 6: Datasets in the VCog-Bench. Both the RAVEN and CVR are rule-based generated datasets. The test samples in MaRs-VQA are designed by psychologists from MaRs-IB.

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.

Table 5: The sub-task statistics in MaRs-VQA.

Sub-task	Proportion
Shape	68%
Colour	73%
Size	16%
Position	41%
Multi-Object	71%

For open-source models, we use the public available weights and data loader settings from the HuggingFace. InstructBLIP (Dai et al., 2024) and MiniGPT-4 (Zhu et al., 2023) are used their original GitHub repo to implement the zero-shot matrix reasoning inference pipeline. Testing is conducted using two NVIDIA RTX 4090 GPUs for 7B-sized VLMs and eight NVIDIA A100 80GB GPUs for VLMs larger than 7B. All experiments are run with three different random seeds, and the results are averaged.

C.2 System Prompts

For each dataset, we prepare custom system prompt. Their pipeline is similar. First, we created a system message prompt (see Figure 6 for zero-shot inference, and Figure 7, 8 for CoT) to guide the MLLM understanding the basic information of matrix reasoning tasks and the structure of the input, and formulating multiple-option images or contexts. The differece for zero-shot and CoT is we provide the guideline to encourage the model think the problem step-by-step based on extracting all useful information from structure $K = \{[r, a, o, s] | r \in \mathcal{R}, a \in \mathcal{A}, o \in \mathcal{O}, s \in \mathcal{S}\}$. The output format is a json structure including "answer" and "reasoning" as keys.

System Prompt for zero-shot inference for MaRs-VQA
You are a helpful visual reasoning assistant that can solve abstract reasoning problems. Each task consisted of a question image with a 3 times 3 matrix. Eight of the nine resulting cells contained an abstract shape, while one cell on the bottom right-hand side of the matrix was empty. Your task is to complete the matrix by finding the missing shape among four possible alternatives. One of the option images is the correct answer. To select the correct missing shape, you have to deduce relationships between the shapes of the matrix. These shape characteristics varied along these dimensions: shape, colour, size, and position in the matrix.

System Prompt for zero-shot inference for RAVEN
You are a helpful visual reasoning assistant solve abstract reasoning problem. Each task consisted of a question image with a 3 times 3 matrix. Eight of the nine resulting cells contained an abstract shape, while one cell on the bottom right-hand side of the matrix was empty. Your task is to complete the matrix by finding the missing shape among eight possible alternatives. One of the option image is the correct answer. To select the correct missing shape, you have to deduce relationships between the shapes of the matrix. These shape characteristics varied along five dimensions: number, shape (triangle, square, pentagon, hexagon, circle), colour (five colours from white to black), size (five size from small to large) and position in the matrix.

Figure 6: System prompts for zero-shot MLLM inference (MaRs-VQA, RAVEN).

D More Discussion on Limitations and Future Work

Limitations. In the main paper, we have briefly discussed the limitations of our work. Here, we provide a more in-depth discussion. First, as our dataset is composed of limited publicly available matrix reasoning datasets (must includes human study results). The RAVEN and CVR datasets, 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. MaRs-VQA can solve this problem, but its samples can not represent all formats of matrix reasoning in the IQ-Tests from Wechsler Intelligence Scale for Children (WISC) and Cattell Culture Faire Intelligence Test (Cattell & Cattell, 1960). We can not use these IQ-Tests directly is because they are not free-to-use, and copyright usually prevents these pen-and-paper tasks from being adapted into computerized tasks.

Second, the size of the datasets in VCog-Bench is relatively small compared with typical computer vision datasets, due to the inherent challenges involved in collecting matrix reasoning data. However, as we claimed in our paper, matrix reasoning should not be presented as typical machine learning settings – finetuning models on training sets and evaluating the performance on test sets. Benchmarking MLLMs' visual reasoning performance should be conducted by zero-shot inference setting, and all data in the test set should not be included in the training set of models. Even compared with other recently released human-designed matrix reasoning datasets, ours is still the largest (see Table 1).

Future work. These limitations highlight the vast potential for future advancements in this field. While our benchmark represents a significant initial step, further data collection and in-depth human studies remain essential. Our experimental results indicate that current MLLMs have enhanced basic matrix reasoning capabilities, with models like GPT-4o and Gemini Pro 1.5 achieving significantly higher accuracy than random guessing across all three matrix reasoning tasks. We anticipate that the next generation of MLLMs will approach human-level performance. It is crucial to maintain the benchmark, continuously monitor the performance of newly released MLLMs, and encourage open-source MLLMs and VLMs to include matrix reasoning tasks for performance comparison.

System Prompt for MLLMs with CoT for MaRs-VQA
You are a helpful visual reasoning assistant that can solve abstract visual reasoning problems. Each task consisted of a question image with a 3 times 3 matrix. Eight of the nine resulting cells contained an abstract shape, while one cell on the bottom right-hand side of the matrix was empty. Your task is to complete the matrix by finding the missing shape among four possible alternatives. One of the options is the correct answer.
The first step is to describe what is the attribute and relationship between each attribute in each cell of the 3 times 3 question image. The attributes can be number, position, shape, size, and color. The cell may contain multiple attributes. The relation might be '3 times 3 sub-blocks', 'rotation', 'insideness'.
The second step is to summarize the relation of three patterns in the first row of the question image, the relation of three patterns in the second row of the question image, the relation of two patterns in the third row of the question image.
Answer this question: What are the row-based high-order rules in the question image?
Based on the description for each option, answer this question: What is the constraint of all options?
Finally, infer what are the potential attributes, objects, relations in the missing cell?

System Prompt for MLLMs with CoT for RAVEN
You are a helpful visual reasoning assistant that can solve abstract visual reasoning problems. Each task consisted of a question image with a 3 times 3 matrix. Eight of the nine resulting cells contained an abstract shape, while one cell on the bottom right-hand side of the matrix was empty. Your task is to complete the matrix by finding the missing shape among eight possible alternatives. One of the option images is the correct answer.
The first step is to summarize the relation of three patterns in the first row of the question image, the relation of three patterns in the second row of the question image, the relation of two patterns in the third row of the question image. What is this relation? The features in the patterns can be constant, progression, arithmetic, distribute three. Try to describe this relationship.
The second step is to describe what is the attribute and relationship between each attribute in each cell of the 3 times 3 cells question image and four option images. The attributes can be number; shape (triangle, square, pentagon, hexagon, circle); colour (five colors: white, light gray, gray, dark gray, black); size (five size: tiny, small, medium, large, huge); and positional relation (inside outside relation, left right relation, top down relation, two times two sub-blocks, 3 times 3 sub-blocks). The cell may contain multiple attributes.
Finally, give me the answer based on step 1-2.

Figure 7: System prompts for CoT MLLM inference (MaRs-VQA, RAVEN).

System Prompt for MLLMs with CoT for CVR
You are a helpful visual reasoning assistant that can solve abstract reasoning problems. Each task consisted of four option images. Your task is to identify which image is different from the other three? To find the correct answer, you have to deduce relationships inside each image and then find the difference. The first step is to describe what is the attribute and relationship between each attribute in four option images. The features can be number of objects; shape; color; size, relationship of colour or shape or size or direction among objects; and positional relation of objects (inside outside relation, left right relation, top bottom relation, adjacent relation). Each image may contain multiple attributes and multiple relations. Based on the description and image for each option, answer this question: What is the constraint / similarity of most of the options? Finally, infer which image is the outlier?

Figure 8: System prompts for CoT MLLM inference (CVR).

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

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 VCog-Bench and 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.