MARVEL: Multidimensional Abstraction and Reasoning through Visual Evaluation and Learning

Yifan Jiang[1*] **Jiarui Zhang**[1*] **Kexuan Sun**[1*] **Zhivar Sourati**[1]
Kian Ahrabian[1] **Kaixin Ma**[2] **Filip Ilievski**[3] **Jay Pujara**[1]
[1]Information Sciences Institute, University of Southern California
[2]Tencent AI Lab, Bellevue, WA
[3]Department of Computer Science, Faculty of Science, Vrije Universiteit Amsterdam
{yjiang44,jzhang37,kexuansu,souratih,ahrabian}@usc.edu
kaixinma@global.tencent.com, f.ilievski@vu.nl, jpujara@isi.edu

Abstract

While multi-modal large language models (MLLMs) have shown significant progress on many popular visual reasoning benchmarks, whether they possess abstract visual reasoning abilities remains an open question. Similar to the Sudoku puzzles, abstract visual reasoning (AVR) problems require finding high-level patterns (e.g., repetition constraints) that control the input shapes (e.g., digits) in a specific task configuration (e.g., matrix). However, existing AVR benchmarks only considered a limited set of patterns (addition, conjunction), input shapes (rectangle, square), and task configurations (3×3 matrices). To evaluate MLLMs' reasoning abilities comprehensively, we introduce **MARVEL**, a multidimensional AVR benchmark with 770 puzzles composed of six core knowledge patterns, geometric and abstract shapes, and five different task configurations. To inspect whether the model accuracy is grounded in perception and reasoning, MARVEL complements the general AVR question with *perception questions* in a hierarchical evaluation framework. We conduct comprehensive experiments on MARVEL with nine representative MLLMs in zero-shot and few-shot settings. Our experiments reveal that all models show near-random performance on the AVR question, with significant performance gaps (40%) compared to humans across all patterns and task configurations. Further analysis of perception questions reveals that MLLMs struggle to comprehend the visual features (near-random performance) and even count the panels in the puzzle (<45%), hindering their ability for abstract reasoning. We release our entire code and dataset.[1]

1 Introduction

Recent advances in novel training pipelines, computational resources, and data sources have enabled Multi-modal Large Language Models (MLLMs) (OpenAI, 2023b; Google, 2023) to show strong visual reasoning ability in tasks that require both visual and textual cues (Wang et al., 2023), such as visual question answering (Goyal et al., 2017a; Antol et al., 2015) and visual commonsense reasoning (Zellers et al., 2019; Xie et al., 2019). These tasks typically focus on testing the models' real-world knowledge (Małkiński, 2023). On the other hand, abstract visual reasoning (AVR) (Zhang et al., 2019; Hill et al., 2019) has little dependency on the world knowledge. As the puzzle shown in Figure 1 (top-right), AVR problems require identifying the hidden pattern (addition and subtraction) that governs the input shapes and their attribute (number of stars/circles) in a task configuration (2×3 matrix). The abstract reasoning ability is related to many practical applications, including visual representations (Patacchiola & Storkey, 2020) and anomaly detection (Schubert et al., 2014),

*Authors contributed equally
[1]https://github.com/1171-jpg/MARVEL_AVR

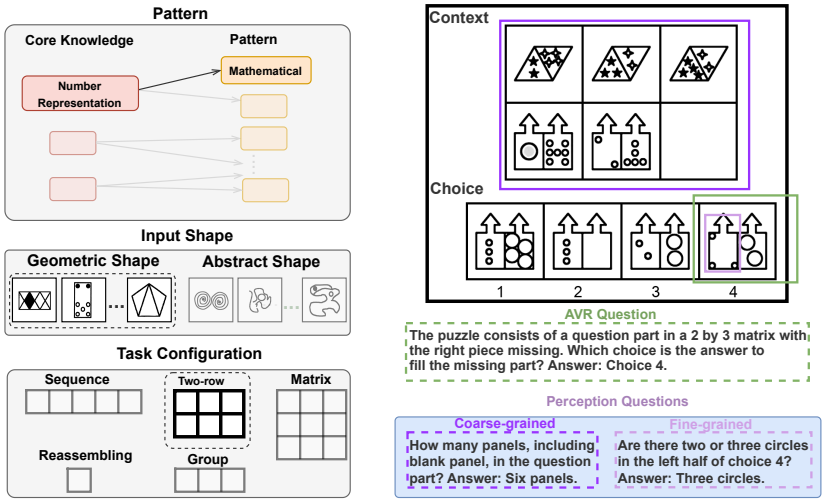


Figure 1: An abstract visual reasoning puzzle in **MARVEL**. The puzzle contains **mathematical** pattern governing the element number in **geometric shapes** with **two-row** task configuration. The AVR question focuses on the final answer for the puzzle, while the perception questions focus on the fine-grained detail about one choice or coarse-grained detail over the whole puzzle.

encouraging more fundamental research in evaluating MLLMs on AVR benchmarks (Ahrabian et al., 2024; Yang et al., 2023b).

However, the scope of current AVR benchmarks is limited, covering few reasoning patterns over a limited set of input shapes arranged in a predetermined configuration of puzzle panels, leading to biased evaluations (van der Maas et al., 2021). RAVEN (Zhang et al., 2019) tests abstract reasoning mostly in mathematical patterns over predefined geometric input shapes. Bongard-LOGO (Nie et al., 2020) and SVRT (Fleuret et al., 2011) implement geometric-related (symmetric) patterns over manually designed abstract input shapes. Although recent surveys highlight the limited evaluation scope of these benchmarks (Małkiński, 2023; van der Maas et al., 2021), a comprehensive benchmark for accessing MLLMs in multidimensional settings has not been proposed and is still needed.

To fill this gap, we introduce **MARVEL**, a multi-dimensional abstract visual reasoning benchmark designed to evaluate MLLMs across different patterns, input shapes, and task configurations. MARVEL's underlying reasoning patterns are rooted in key core knowledge of human cognition, observed in newborn infants relying only on abstraction to reason about their environment (Spelke & Kinzler, 2007). To facilitate a more comprehensive evaluation of foundation MLLMs, MARVEL combines six patterns expanded from three types of core knowledge, both geometric and abstract shapes, and five task configurations. We crawl relevant puzzles from publicly available websites, manually filter low-quality and irrelevant puzzles based on the expanded patterns and input shapes, and reformat them into different task configurations. We annotate the AVR question by providing a short description of the puzzle and asking for its answer. In total, we collect 770 diverse and high-quality puzzles assessing abstract visual reasoning abilities (Figure 1).

To provide a hierarchical evaluation, we also enrich each puzzle with perception questions focusing on perceiving puzzles' visual details (e.g., number of grids, edges of a triangle) to measure models' reasoning consistency (Jiang et al., 2023; Selvaraju et al., 2020). We conduct comprehensive experiments on MARVEL involving different model structures, model sizes, and prompting strategies. Our experiments reveal that all MLLMs show near-random performance in all patterns, even with few-shot demonstrations and prompt engineering, leaving a huge gap (40%) in the abstract reasoning ability with humans. An in-depth

Dimension		RAVEN	G-set	VAP	Bongard-LOGO	SVRT	DOPT	ARC*	**MARVEL**
Input Shape	Geometric	✔	✔	✔			✔	✔	✔
	Abstract				✔	✔			✔
Pattern	Temporal Movement	✔	✔				✔	✔	✔
	Spatial Relationship					✔		✔	✔
	Quantities	✔	✔	✔	✔	✔	✔	✔	✔
	Mathematical	✔	✔	✔				✔	✔
	2D-Geometry				✔	✔			✔
	3D-Geometry								✔
Configuration	Sequence						✔		✔
	Two-row			✔					✔
	Matrix	✔	✔						✔
	Group				✔	✔			✔
	Reassembling								✔
Perception Question									✔

Table 1: Comparing MARVEL to related benchmarks: RAVEN (Zhang et al., 2019), G-set (Mańdziuk & Żychowski, 2019), VAP (Hill et al., 2019), Bongard-LOGO (Nie et al., 2020), SVRT (Fleuret et al., 2011), ARC (Chollet, 2019), DOPT (Webb et al., 2020). *ARC puzzles are provided in a generative format.

analysis based on perception questions points out that MLLMs' poor reasoning ability is hindered by their struggle with fine-grained visual feature comprehension. Their near-random perception ability fails to provide foundations for subsequent abstract reasoning. In summary, our contributions can be listed as follows: 1) **A novel multidimensional AVR benchmark**, **MARVEL**, which consists of six patterns across five distinct task configurations. 2) **A hierarchical evaluation framework** incorporating perception questions with AVR questions to enable fine-grained diagnosis of model capability. 3) **Extensive experiments** on a wide range of state-of-the-art MLLMs with various prompting strategies, providing insights into their strengths, weaknesses, and future improvement directions on AVR tasks and MLLM developments.

2 Related Work

MLLM Evaluations. Benefiting from the rich representation from visual encoders (Radford et al., 2021) and strong reasoning ability of LLMs (Touvron et al., 2023; Chiang et al., 2023), MLLMs (Li et al., 2023; Dai et al., 2024; OpenAI, 2023a; Liu et al., 2024) have been applied to solve not only traditional vision-language tasks, such as image captioning (Agrawal et al., 2019; Young et al., 2014), visual question answering (Goyal et al., 2017b; Marino et al., 2019; Hudson & Manning, 2019a; Singh et al., 2019) and refer expression comprehension (Kazemzadeh et al., 2014; Gupta et al., 2022), but also on more complicated scenarios, such visually-grounded conversation (Liu et al., 2024; Alayrac et al., 2022), multimodal web/UI agents (He et al., 2024; Zheng et al., 2024; Yang et al., 2023a) and embodied tasks Driess et al. (2023). Besides end-to-end evaluation, several recent works also try to reveal MLLMs' visual shortcomings from different aspects, including visual details (Zhang et al., 2023; Wu & Xie, 2023), perceptual bias (Zhang et al., 2024), and small visual pattern recognition (Tong et al., 2024). Although some of the existing benchmarks have accessed MLLM's mathematical visual reasoning abilities requiring an understanding of abstract and geometry shapes (Lu et al., 2023; 2022), their evaluation still heavily relies on textual descriptions. In contrast, AVR benchmarks assess MLLMs' ability under a diverse set of patterns with only visual understanding settings.

AVR Benchmarks. AVR problems have great potential impact on various domains (Małkiński, 2023; Patacchiola & Storkey, 2020; Schubert et al., 2014), sparking interests in evaluating MLLMs on AVR benchmarks (Ahrabian et al., 2024; Moskvichev et al., 2023; Mitchell et al., 2023). Existing AVR benchmarks present the evaluation in a wide range of formats, such as selective completion (Zhang et al., 2019; Hu et al., 2021; Benny et al., 2021; Webb et al., 2020), group discrimination (Fleuret et al., 2011; Nie et al., 2020) and generative completion (Chollet, 2019). However, less attention is paid to the scope and

pattern of the AVR benchmark; most focus only on a few simple abstract patterns and testing models end-to-end without considering the intermediate perception and reasoning procedures (Moskvichev et al., 2023; Mitchell, 2021). In contrast, MARVEL includes geometric and abstract shapes, six core patterns essential for visual abstract reasoning, and five different task configurations. Inspired by prior analysis of visual details and perceptual bias, MARVEL introduces perception questions to ensure the MLLMs correctly perceive the presented visual patterns. Our work and other related AVR benchmarks are compared in Table 1.

3 MARVEL Benchmark Construction

As a multidimensional benchmark for AVR, MARVEL covers different task configurations (Section 3.1), various input shapes (Section 3.2), as well as different reasoning patterns involved in the puzzles (Section 3.3). Further, MARVEL decomposes the evaluation of models' capabilities into 1) perception questions (Section 4) about the puzzle panels and 2) AVR questions (Section 3.4) to yield a more precise and faithful picture of evaluated models. The data collection process is presented in Section 3.4.

3.1 Task Definition and Configurations

Each puzzle in MARVEL consists of a context on the top and possible choices ($c_i; i \in \{1, 2, 3, 4\}$) to choose from at the bottom, formatted in a multiple-choice question answering setting (see Figure 1 for an example). The context part consists of n puzzle panels ($p_1, p_2, \ldots, p_n, p_b$ with p_b being a blank panel), with their specific number and arrangement driven by a task configuration and reasoning pattern, P, that governs the relationship between puzzle panels. The choice will be considered the correct answer and fill in p_b that can satisfy the following equation: $P(p_1, p_2, \ldots, p_n) = P(p_1, p_2, \ldots, p_n, c_c)$.

Puzzle panels in MARVEL are organized in the following five task configurations:

1. **Sequence Format** arranges panels in a $1 \times n$ line ($n \in [4, 7]$).
2. **Two-row Format** presents panels in a 2×3 matrix*, p_1^1, p_2^1, p_3^1 and p_1^2, p_2^2, p_3^2. The solution requires identifying the same pattern at the first row and the second row, which is $P(p_1^1, p_2^1, p_3^1) = P(p_1^2, p_2^2, c_c)$.
3. **Matrix Format** organizes panels in a 3 by 3 matrix, the pattern can be reflected in either row- or column-wise way: $P(p_1^1, p_2^1, p_3^1) = P(p_1^2, p_2^2, p_3^2) = P(p_1^3, p_2^3, c_c)$ or $P(p_1^1, p_1^2, p_1^3) = P(p_2^1, p_2^2, p_3^2) = P(p_3^1, p_3^2, c_c)$.
4. **Group Format** has three panels in the question part (p_1, p_2, p_b) with one choice reflecting the context's pattern and other choices differing: $P(p_1, p_2, c_c) \neq P'(choices - c_c)$.
5. **Reassembling Format** is designed for *3D-Geometry* pattern with one unfolded diagram in the context and four 3D assembly results as choices.

We visualize these configurations in Figure 1 (see detailed examples in Appendix A).

3.2 Input Shapes

As shown in Figure 1, each panel of a puzzle contains various shapes that can be generally differentiated into two types (Małkiński, 2023):

1. **Geometric Shapes** are easily comprehended and described. For instance, a square is a shape that has four sides of equal length and four equal angles. Most existing AVR benchmarks (Zhang et al., 2019; Hill et al., 2019) focus on elementary shapes such as oval, rectangle, triangle and trapezoid. MARVEL includes geometric shapes consisting of more than two different elementary geometric shapes to mitigate the issue and improve the complexity.
2. **Abstract Shapes** come from a wide set of possibilities and vary widely from one problem to another. It provides a fair step as most MLLMs encounter the shapes for the first

*p_a^b: the panel on the a th row and b th column of matrix.

time. MARVEL also includes abstract shapes as they are gaining more preference for AVR-related research (Fleuret et al., 2011; Nie et al., 2020).

3.3 Core Knowledge and Patterns

Core knowledge theory (Spelke & Kinzler, 2007) from cognition developmental psychology is largely shared among humans and particularly for human infants. Human infants with no real-world knowledge and limited experience represent their environment using abstraction patterns. These abstraction patterns can be categorized into three types of core knowledge, which is the foundation for inference and reasoning (Lipton & Spelke, 2004). We expand each core knowledge[2] into two patterns for a fine-grained assessment of abstract reasoning in MARVEL, based on insights drawn from contemporary cognitive literature:

Object Core Knowledge represents objects' spatio-temporal motions and their contact, which is expanded to *Temporal Movement Pattern* focusing on the related position change or movement (Spelke, 1990) and *Spatial Relationship Pattern* examining objects' relative positional relationship (Aguiar & Baillargeon, 1999).

Number Core Knowledge helps infants process abstract representations of small numbers. We include *Quantities Pattern* testing the accuracy of number comprehension (Xu & Spelke, 2000) and *Mathmatical Pattern* for elementary mathematical operations (Barth et al., 2005).

Geometry Core Knowledge captures the environment's geometry, which helps humans orient themselves in their surroundings. We divide it into *2D-Geometry Pattern* (Hermer & Spelke, 1996) and *3D-Geometry Pattern* (Cheng & Newcombe, 2005).

3.4 Data Collection

We collect puzzles from several public resources websites[3] and filter out unfit or low-quality data by three human annotators based on the puzzle's input shapes (some puzzles contain textual information) and patterns. Unaligned puzzles are first segmented into panels and then reassembled into the correct task configuration. To ensure each pattern in each task configuration has at least 45 puzzles[4], we also manually created 220 puzzles by following the pattern in existing data and replacing the input shape drawn from scratch. Each puzzle contains an AVR question (Figure 1) generated from templates based on their task configuration. AVR questions provide a brief description and ask only for the puzzle's final answer, which is widely adopted in previous AVR benchmark (Małkiński, 2023). In the end, MARVEL includes 770 high-quality puzzles over six high-level patterns across five distinct task configurations. In Table 1, we compare MARVEL with existing AVR benchmarks to show its comprehensive scope.

4 Hierarchical Evaluation Framework

Previous works evaluate MLLMs on AVR benchmarks with the final answer only (Moskvichev et al., 2023; Mitchell et al., 2023), potentially overlooking shortcut learning and inductive biases (Małkiński, 2023). On the other hand, precisely comprehending visual details is the foundation for subsequent reasoning in AVR problems (Gao et al., 2023). We enrich MARVEL puzzles with perception questions (Selvaraju et al., 2020) designed to test models' perception ability on visual details (Figure 1). We design a hierarchical evaluation framework by combining two types of perception questions with AVR questions (Figure 1) to examine if model accuracy is based on perception and reasoning. For each puzzle, our framework provides three coarse-grained questions and one pattern-related fine-grained question:

[2]We don't feature the fourth Core Knowledge, agent representation, due to its concentrates on goal-directed and interactive action, which is not adaptable in MARVEL setting.

[3]`https://www.gwy.com/`; `https://www.chinagwy.org/`

[4]Some patterns can not be presented in specific configurations. For example, the *Mathmatical Pattern* can not be adapted to group format as it requires comparison between adjacent panels.

Coarse-grained Perception Question in an open-ended fashion aims to test if models can understand task configuration correctly, which directly asks about the number of panels in puzzles. We use templates to generate three questions focusing on the number of panels in the context part, choice parts, and the whole puzzle. We remove the choice index when testing models with this question to avoid shortcut learning.

Fine-grained Perception Question in binary-choice format examines models' understanding of input shapes, which focus on the visual details (Selvaraju et al., 2020) such as shape attributes (number of edges) and spatial relationship (left, right) based on the pattern contained in the puzzle. For example, in Figure 1, the fine-grained perception question tests whether models can understand the number of circles because the puzzle contains *Mathematical Pattern*. For each puzzle, we randomly pick one choice panel in the puzzle and manually create questions with two choices. The correct answer is randomly placed to avoid inductive bias. We have five types of questions based on the pattern and how it adapts to the input shape, which are listed with examples here:

1. **Location**: Is the dot outside or inside of the star in choice 4?
2. **Color**: Is the triangle black or white in choice 1?
3. **Shape**: Is there a circle or a triangle inside choice 3?
4. **Quantity**: Are there five or four circles in choice 2?
5. **Comparison**: Are the left and right halves of the rectangle in choice 3 the same?

5 Experimental Setup

5.1 Model Selection

Closed-source MLLMs. We include API-based MLLMs including 1) GPT-4V (OpenAI, 2023a), 2) Gemini (Google, 2023) and 3) Claude3 (Anthropic, 2024). With the massive computation and training data, these models show promising performance on a wide range of visual-focused tasks (Goyal et al., 2017b; Xie et al., 2019). We evaluate closed-source MLLMs in both zero-shot and few-shot (Brown et al., 2020) settings.

Open-source MLLMs. We include MLLMs smaller than 13B due to our limited computing resources: 1) InstructBLIP (Dai et al., 2024), 2) BLIP-2 (Li et al., 2023), 3) Fuyu (Bavishi et al., 2023), 4) Qwen-VL (Bai et al., 2023) and 5) LLaVA (Liu et al., 2023). We only evaluate these MLLMs in a zero-shot setting due to their single-image input settings (Zhao et al., 2023).

Human Evaluation. To access the upper bound performance on MARVEL, we simulate a realistic human assessment by inviting 30 annotators aged from 10 to 50 years to solve a subset of MARVEL and ensure each subset contains every pattern in all task configurations. We compute the average performance of these 30 annotators as the human baseline. Each puzzle is solved by at least two annotators.

5.2 Evaluation Metrics

Following a similar setting as previous research evaluating MLLMs on the AVR benchmark (Ahrabian et al., 2024), we use regex matching to extract the choices picked (e.g., "choice 4" in the response "The correct answer is choice 4."), with failure cases re-extracted by GPT-4 (Aher et al., 2023). We use accuracy as the metric, which is commonly used for evaluating multiple-choice questions and has been utilized by many AVR systems (Zhang et al., 2019; Hill et al., 2019). Based on the hierarchical evaluation framework, we evaluate MLLMs with two types of accuracy-based metrics:

1. **Instance-based Accuracy** considers questions separately. We report accuracy results for *AVR question* and *fine-grained perception question*.
2. **Group-based Accuracy** considers questions as group to assess the consistency in model reasoning (Jiang et al., 2023; Yuan et al., 2021). The model receives a score of 1 only if it correctly answers all questions within the same group. We report the group-based

accuracy result of combining all three coarse-grained perception questions and the further result after introducing fine-grained and AVR questions into the group.

Category	Model	AVR Question	Fine-grained Perception	Perc^{C}	$\text{Perc}^{\text{C\&F}}$	$\text{Perc}^{\text{C\&F}}$ & AVR
Random		25.00	50.00	-	-	-
Open-source MLLMs	Qwen-VL (7B)	19.61	37.27	0.52	0.39	0.00
	Fuyu (8B)	24.29†	34.94	0.00	0.00	0.00
	BLIP-2 (FlanT5$_{\text{XXL}}$-11B)	24.81†	**53.38**	1.04	0.52	0.26
	InstructBLIP (Vicuna-13B)	24.68†	49.48	0.00	0.00	0.00
	LLaVA-1.5 (Vicuna-13B)	26.36	51.43	1.14	0.52	0.13
Closed-source MLLMs	GPT-4V	22.34	51.56	18.31	9.22	2.73
	Gemini-pro-vision*	25.06†	44.42	15.19	6.75	1.69
	Claude3 (Sonnet)	26.49†	50.91	38.70	19.87	5.06
	Claude3 (Opus)	**28.83**	47.27	**44.94**	**20.13**	**5.97**
Human		68.86 ± 9.74	-	-	-	-

Table 2: Main zero-shot accuracy over MARVEL across all MLLMs in two accuracy metrics: $Prec^{\text{C}}$ = group-based accuracy over all coarse-grained perception questions (model must answer all three questions correctly), $Perc^{C\&F}$ = group-based accuracy combining all perception questions (coarse/fine-grained), AVR = AVR Question. The best performance among all models is in **bold**, and the best result in two MLLMs categories is underlined. *Gemini refuses to answer the puzzle due to safety problems in 7% cases so the performance is computed based on the left set. † notes the result is attributed to inductive bias.

6 Results

We focus on four research questions: 1) What's the abstract reasoning ability on visual puzzles of current SOTA MLLMs? 2) Can MLLMs do better with different few-shot prompting strategies? 3) How do MLLMs perform on different patterns and task configurations? 4) To what extent do MLLMs visually understand the puzzle, and do they show consistent reasoning ability?

Overall Performance. The AVR question results are shown in Table 2. Human performance reaches 68.86%, with a standard deviation of 9.74, confirming the validity and challenging nature of MARVEL. For both open and closed source categories, **all models show near-random performance with a huge gap (40%) compared to human performance**, in which closed-source MLLMs (avg: 25.7%) perform slightly better than open-sourced ones (avg: 24.0%). We observed an extremely imbalanced distribution in the outputs of some MLLMs. For example, BLIP2 consistently selecting choice 1 for all puzzles (marked † in Table 2). We tried different approaches with our best effort to avoid potential bad prompts or engineering settings, including adding question marks in the black panel, replacing the choice index with letter (1 → A), and changing the description in the AVR question. None of them can mitigate and may even exacerbate the issue, highlighting the potential inductive biases (Wang & Wu, 2024) in models. Among open-source MLLMs, LLAVA performs the best, yet the gap is very small, and it is unclear whether the gain comes from its larger model size. In closed-source models, even the strongest MLLM, Claude3 (Opus), which demonstrated promising results on various vision tasks (Anthropic, 2024), failed to present a significant performance difference from the random baseline. Claude3 (Sonnet) and Gemini also have imbalanced output distributions, with both selecting choice 4 in most cases.

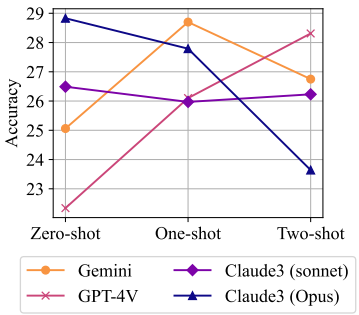
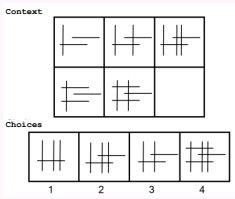


Figure 2: MLLMs performance in different few-shot COT.

Impact of Few-Shot CoT. Given the poor zero-shot performance and models' capability of in-context learning (Brown et al., 2020), we explore few-shot prompting with Chain-of-Thought (CoT) (Wei et al., 2022) to guide MLLMs with abstract reasoning patterns.

We experiment with all closed-source MLLMs in one-shot and two-shot settings using manually created CoT context, similar to Yang et al. (2023b). For each puzzle, we randomly select puzzles with the same pattern in the sequence task configuration and annotate the CoT reasoning with answers as demonstrations. We chose the sequence task because it is more straightforward (only along the sequence) compared to other configurations. Each demonstration is formatted as image-text pairs. Our result is shown in Figure 2, and we present the full results in Appendix E. The few-shot demonstrations show a marginal positive impact on GPT-4V and a decreasing trend on Claude3 (Opus).

Further analysis reveals that the main improvement in GPT-4V's results lies in the *3D-Geometry pattern*. As this pattern focuses on reassembling, the demonstration can guide the model to pay attention to the relative position of each face of the object. However, since most patterns are uniquely implemented on different input shapes and their attributes, the model struggles to learn generalizable patterns from the few-shot demonstrations. Figure 3 provides an example of zero-and few-shot results from Claude3 (Opus). With the demonstration, the model learns to focus on the correct pattern (blue) at the beginning of the reasoning. However, it fails to adapt precisely to the input shapes in the puzzle (red), leading to errors in subsequent reasoning. We also test different prompt engineering approaches, including selecting demonstration samples from different 1) patterns, 2) task configurations, and 3) prompting MLLMs by dividing puzzles panel by panel. None of these approaches yields a positive impact; instead, they lead to a significant drop in performance (AppendixE). Given the complexity and challenging nature of the dataset, the effectiveness of few-shot prompting on MARVEL remains minimal.[5]



Perception Question:
Are there two or three parallel horizontal lines in choice 2?

Claude3-Opus:
Choice 2 contains three parallel horizontal lines.

Claude3-Opus (Zero-shot):
To determine which choice best fits the missing piece in the bottom right, we need to identify the pattern in the given 3x2 grid. Looking at the grid, we see that each row follows a specific pattern:

Row 1: The number of vertical lines increases by 1 from left to right.
Row 2: The number of horizontal lines increases by 1 from left to right.
Based on this pattern, the missing piece in the bottom right should have 3 horizontal lines, as it follows the pattern in the second row.

Looking at the choices provided, choice 3 is the only option that has 3 horizontal lines. Therefore, **choice 3** is the most appropriate answer to fill the missing part in the given puzzle.

Claude3-Opus (Two-shot):
Let's analyze the pattern in the given 3x2 matrix, focusing on the number of vertical and horizontal lines in each cell

Row 1: Cell 1 has one vertical line, Cell 2 has two vertical lines
Row 2: Cell 1 has one vertical and one horizontal line, forming a plus sign. Cell 2 has two vertical and two horizontal lines, forming a hash symbol.
Row 3: Cell 1 has three vertical lines. The missing cell should follow the pattern established in the previous rows.

Based on the pattern, the missing cell should have three vertical and three horizontal lines, forming a grid of nine squares. This corresponds to choice 4 among the given options. Therefore, **choice 4** is the most appropriate answer to complete the matrix.

Figure 3: Perception question, zero- and two-shot example of Claude3 (Opus).

Performance on Different Patterns and Configurations. We further break down the results[6] into different patterns and task configurations in Figure 4 (full results in Appendix B). In general, MLLMs show near-random performance on all patterns and task configurations. Among the six patterns, *3D-Geomertry* pattern is the most challenging for humans and MLLMs. The difficulty may be rooted in the requirement for the 3D imaginative ability (Margules & Gallistel, 1988) and relatively less common in the models' pre-training datasets, which leads to a significant gap (12.5%) between open and closed-sources MLLMs. On the other hand, *2D-Geomertry*, involving understanding geometric attribute of input shape, seems relatively easier for MLLMs, evidenced by their proficiency on similar tasks such as interpreting graphs plots (Yang et al., 2023b). In two model categories, LLaVA ranks first on three patterns, showing competitive performance compared to closed-source MLLMs. Claude3 (Opus) exhibits a balanced and strong reasoning ability, ranking in the top 2 across all patterns.

Among the five task configurations, the two-row and matrix formats provide more panels (i.e., information) to verify the pattern, facilitating puzzle-solving. On the contrary, the group and reassembling formats containing the least panel number tend to be challenging.

[5]We find the inductive bias is mitigated in few-shot settings.
[6]We only select MLLMs that do not show significantly imbalanced output distributions.

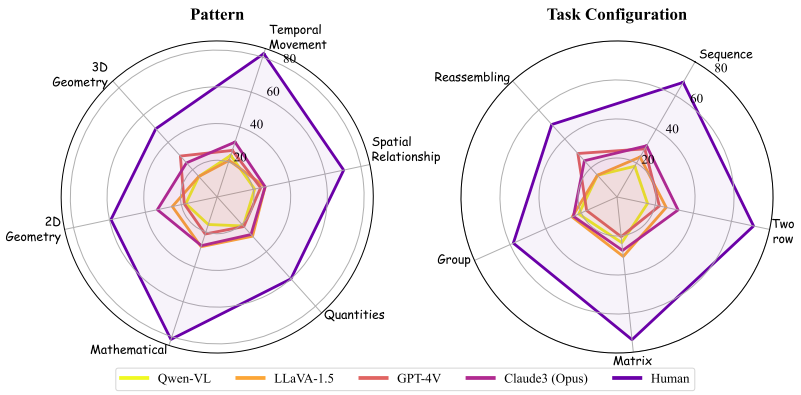


Figure 4: MLLMs and human performance across patterns and task configurations.

Three out of four MLLMs rank 1st in different task configurations, which verifies our assumption of potential bias in single-configuration evaluation. Models may be familiar with specific input types according to their pre-training dataset, highlighting the importance and necessity of multidimensional settings in MARVEL.

Perception Ability and Reasoning Consistency. Visual cognition forms the foundation for advanced reasoning (Richards et al., 1984). By incorporating perception questions, our hierarchical evaluation framework effectively investigates to what extent the models understand the visual information from the puzzle. In Table 2, closed-source MLLMs demonstrate more robust performance on coarse-grained perception group accuracy compared to open-sourced MLLMs, with a performance gap ranging from 14.05% to 44.80%. However, even the best model fails to reach 50% accuracy, indicating that current MLLMs struggle to simultaneously understand the number of grids, choices, and the puzzle as a whole, despite their promising performance on real-world datasets (Hudson & Manning, 2019b). The simplicity of the coarse-grained perception questions (all puzzles contain less than 13 panels) highlights the poor perception ability of current MLLMs in the abstract visual reasoning domain. Fine-grained perception questions further confirm this argument, with all models showing near-random performance. Further analysis of fine-grained perception performance based on five categories (Table in Appendix C) reveals that models perform relatively better at color perception but have difficulty recognizing location (e.g., 'a' is on the left of 'b'). We hypothesize that the difficulty in understanding location stems from the lack of labeled data on location and relations during training, especially in abstract visual understanding. In contrast, the models' color perception is well-trained during their multi-modal alignment, and the simplicity of RGB understanding allows for easier transfer to the abstract domain.

The further group-based accuracy ($Prec^{C\&F}$ and $Prec^{C\&F}\&AVR$) shows that **no model can solve the AVR puzzles with consistent reasoning**, with the best model reaching only 5.97% group accuracy. Based on the result of our evaluation framework, we hypothesize the inconsistency stems from their poor visual perception ability (Gao et al., 2023). To verify our assumption, we run an additional experiment on a subset of MARVEL by adding accurate text descriptions of the puzzle in the input (full result in Appendix D). The result shows a significant boost in performance (11.57% to 44.21%), with GPT-4V even displays on-par performance (65.26%) with humans. As shown in Figure 3, the model's reasoning is based on the perception of the puzzle (e.g., number of lines), which needs to be completely precise to support correct reasoning. The perception questions in our framework reveal that the model cannot clearly understand the number of lines, explaining why it fails to answer the puzzle even with correct hints (few-shot). A single error in visual feature perception can impact reasoning since the correct pattern must apply to all puzzle shapes. The densely packed information distribution—where the majority of the puzzle remains blank—ensures that each piece of visual perception is an essential foundation for subsequent reasoning. However, the importance of visual detail perception has received little attention in previous evaluations (Moskvichev et al., 2023; Mitchell et al., 2023), highlighting the significance of our new evaluation benchmark.

7 Conclusion

In this work, we developed MARVEL, a multidimensional abstract reasoning benchmark comprising 770 puzzles with both geometric and abstract input shapes across six patterns and five task configurations. We also designed a hierarchical evaluation framework that enriches MARVEL with perception questions to enable granular analysis of models' visual details understanding and reasoning consistency. Our comprehensive experiments with nine SoTA MLLMs revealed a huge gap in abstract visual reasoning ability between (40%) humans and MLLMs, where all MLLM often perform close to random. Further analysis based on our evaluation framework showed MLLMs' poor perception ability in understanding visual details, which hinders their subsequent reasoning and leads to poor AVR performance. We hope future works can build on the foundation of MARVEL for enhancing MLLM abstract visual perception and reasoning abilities.

8 Acknowledgements

We appreciate Fred Morstatter for very helpful comments. We thank Tian Jin and Jiachi Liang for their assistance in data collection. This research was sponsored by the Defense Advanced Research Projects Agency via Contract HR00112390061.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

Andréa Aguiar and Renée Baillargeon. 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive psychology*, 39(2):116–157, 1999.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pp. 337–371. PMLR, 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.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anthropic. Claude 3, 2024. URL `https://www.anthropic.com/news/claude-3-family`.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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.

Hilary Barth, Kristen La Mont, Jennifer Lipton, and Elizabeth S Spelke. Abstract number and arithmetic in preschool children. *Proceedings of the national academy of sciences*, 102(39): 14116–14121, 2005.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL `https://www.adept.ai/blog/fuyu-8b`.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ken Cheng and Nora S Newcombe. Is there a geometric module for spatial orientation? squaring theory and evidence. *Psychonomic bulletin & review*, 12:1–23, 2005.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

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.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

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.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

Gemini Team Google. Gemini: A family of highly capable multimodal models, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017a.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017b.

Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: General robust image task benchmark. *arXiv preprint arXiv:2204.13653*, 2022.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models, 2024.

Linda Hermer and Elizabeth Spelke. Modularity and development: The case of spatial reorientation. *Cognition*, 61(3):195–232, 1996.

Felix Hill, Adam Santoro, David GT Barrett, Ari S Morcos, and Timothy Lillicrap. Learning to make analogies by contrasting abstract relational structure. *arXiv preprint arXiv:1902.00120*, 2019.

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.

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

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

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. Brainteaser: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14317–14332, 2023.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

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

Jennifer S Lipton and Elizabeth S Spelke. Discrimination of large and small numerosities by human infants. *Infancy*, 5(3):271–290, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

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

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.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*, 2023.

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

Jacek Mańdziuk and Adam Żychowski. Deepiq: A human-inspired ai system for solving iq test problems. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Joan Margules and CR Gallistel. Heading in the rat: Determination by environmental shape. *Animal Learning & Behavior*, 16(4):404–410, 1988.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

Melanie Mitchell, Alessandro B Palmarini, and Arsenii Kirillovich Moskvichev. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. In *AAAI 2024 Workshop on"Are Large Language Models Simply Causal Parrots?"*, 2023.

Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*, 2023.

Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. *Advances in Neural Information Processing Systems*, 33:16468–16480, 2020.

OpenAI. Gpt-4 technical report, 2023a.

R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13, 2023b.

Massimiliano Patacchiola and Amos J Storkey. Self-supervised relational reasoning for representation learning. *Advances in Neural Information Processing Systems*, 33:4003–4014, 2020.

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.

WA Richards et al. Parts of recognition. *Cognition*, 18(1), 1984.

Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data mining and knowledge discovery*, 28:190–237, 2014.

Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10003–10011, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1): 89–96, 2007.

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

Han LJ van der Maas, Lukas Snoek, and Claire E Stevenson. How much intelligence is there in artificial intelligence? a 2020 update. *Intelligence*, 87:101548, 2021.

Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *Meta-Radiology*, pp. 100047, 2023.

Zhenhailong Wang, Joy Hsu, Xingyao Wang, Kuan-Hao Huang, Manling Li, Jiajun Wu, and Heng Ji. Text-based reasoning about vector graphics. *arXiv preprint arXiv:2404.06479*, 2024.

Zihao Wang and Lei Wu. Theoretical analysis of the inductive biases in deep convolutional networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Taylor Webb, Zachary Dulberg, Steven Frankland, Alexander Petrov, Randall O'Reilly, and Jonathan Cohen. Learning representations that support extrapolation. In *International conference on machine learning*, pp. 10136–10146. PMLR, 2020.

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

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Fei Xu and Elizabeth S Spelke. Large number discrimination in 6-month-old infants. *Cognition*, 74(1):B1–B11, 2000.

Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023a.

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, 2023b.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, and Tsong Yueh Chen. Perception matters: Detecting perception failures of vqa models using metamorphic testing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16908–16917, 2021.

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.

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.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Visual cropping improves zero-shot question answering of multimodal large language models. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. Exploring perceptual limitation of multimodal large language models. *arXiv preprint arXiv:2402.07384*, 2024.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded, 2024.

A Data Examples

We present examples with 5 different task configurations and 6 patterns in Figures 5 to 11.

B Performance on Different Pattern and Tasks

Table 3 and Table 4 shows MLLMs' AVR question performance on different patterns and tasks.

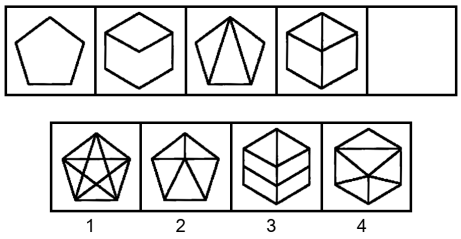


Figure 5: The example is formatted in *Sequence* configuration with the *Quantities* pattern. The answer to this puzzle is B.

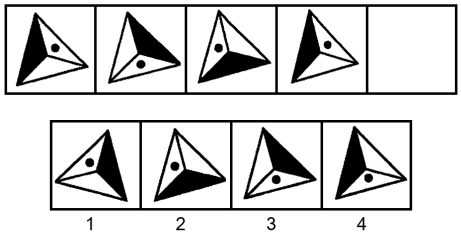


Figure 6: The example is formatted in *Sequence* configuration with the *Temporal Movement* pattern. The answer to this puzzle is C.

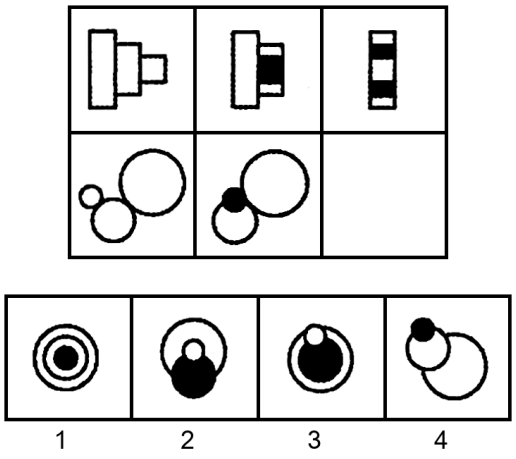


Figure 7: The example is formatted in *Two-row* configuration with the *Spatial Relationship* pattern. The answer to this puzzle is B.

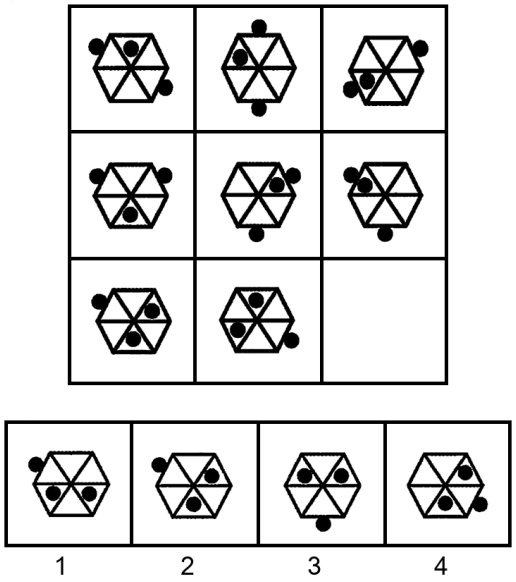


Figure 8: The example is formatted in *Matrix* configuration with the *Spatial Relationship* pattern. The answer to this puzzle is B.

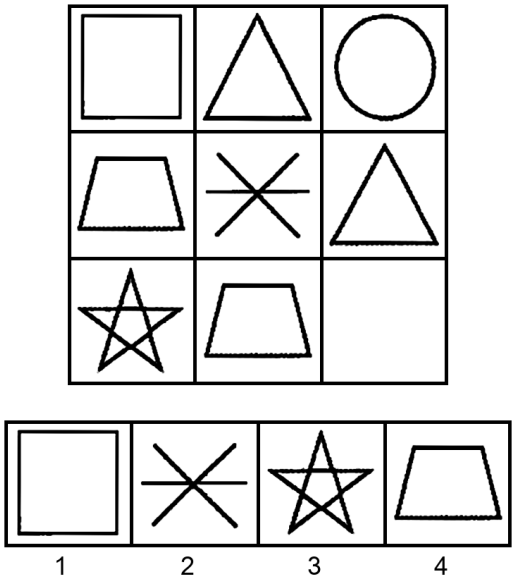


Figure 9: The example is formatted in *Matrix* configuration with the *Mathematical* pattern in *Geometric* shapes. The answer to this puzzle is A.

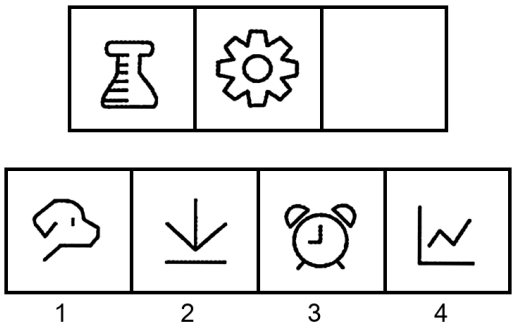


Figure 10: The example is formatted in *Group* configuration with the *2D-Geometry* pattern in *Abstract* input shapes. The answer to this puzzle is C.

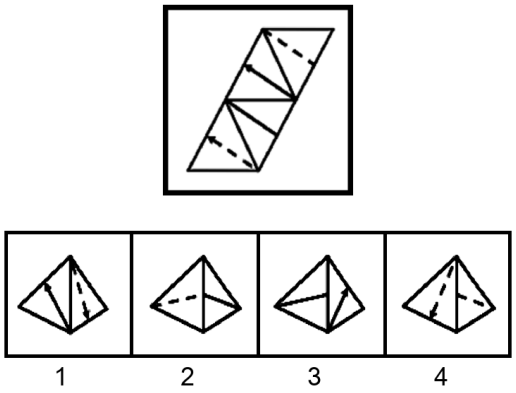


Figure 11: The example is formatted in *Reassembling* configuration with the *3D-Geometry* Quantities. The answer to this puzzle is B.

C Coarse-grained Perception on Different Categories

Table 5 shows MLLMs coarse-grained perception performance on different categories.

D Question with text description

To investigate how the models perform subsequent reasoning with accurate visual detail awareness, we randomly select 95 puzzles (five for each pattern in each task configuration) and provide text descriptions for each panel in the puzzle. To ensure a similar level of granularity for text descriptions, we first use GPT-4V to provide the original text descriptions, which human annotators further modify without introducing unrelated details (Figure 12).

We input MLLMs with both the AVR question and text descriptions. The result is shown in Table 6. With the help of text descriptions, MLLMs, especially closed-sourced MLLMs, can build their abstract visual reasoning on correct visual detail foundations, gaining significant improvement in the performance (11.57% to 44.21%). GPT-4V even shows on-par performance (65.26%) with humans, highlighting the importance of visual perception ability. We also want to point out that some puzzles containing abstract shapes are challenging to

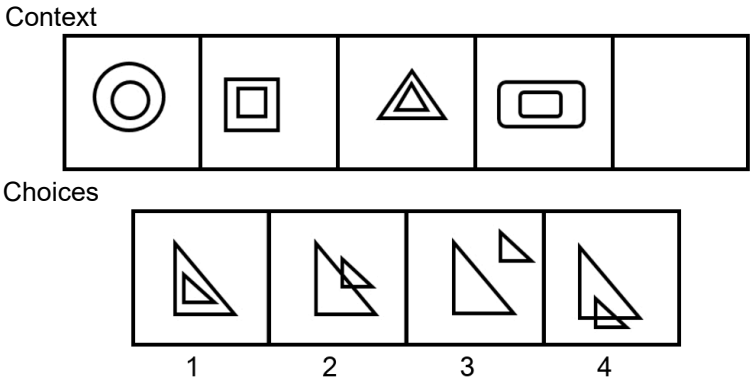


Figure 12: An example of the annotation process. We first generate text descriptions from GPT-4V, and the output will further be modified by human annotators (red → green).

Pattern	Human	Open-sourced MLLMs					Closed-sourced MLLMs			
		Qwen-VL	Fuyu†	Blip-2†	InstructBLIP†	LLaVA-1.5	GPT-4V	Gemini†	Claude3 (Sonnet)†	Claude3 (Opus)
Temporal Movement	82.08	23.81	24.76	25.71	25.71	20.95	26.67	22.86	23.81	31.43
Spatial Relationship	70.42	20.83	25.00	26.67	26.67	26.67	24.17	29.17	34.17	26.67
Quantities	81.57	15.76	24.85	27.88	27.27	28.48	21.21	24.85	24.85	27.88
Mathematical	60.00	21.25	25.00	24.17	24.17	28.75	21.67	23.75	23.33	27.50
2D-Geometry	59.17	17.50	22.50	20.83	20.83	25.00	18.33	25.00	30.83	33.33
3D-Geometry	50.00	15.00	15.00	15.00	15.00	15.00	30.00	30.00	20.00	25.00

Table 3: Performance of different models for different patterns. † notes the result is attributed to inductive bias.

describe. A tool that can convert images to SVGs and text descriptions (Wang et al., 2024) can be a possible approach to mitigate the difficulty and enhance MLLMs performance.

E Few-Shot COT

We show our prompting template in Table 11. We discuss few-shot performance in Tables 7 to 10. Table 7 shows few-shot result with Chain-of-Thought demonstrations in the prompt. Table 8 compares one-shot result with Chain-of-Thought demonstrations from same pattern or different pattern picked randomly (OOD). Table 9 compares one-shot result with Chain-of-Thought demonstrations from single task format (sequence) and two different task

Task format	Human	Open-sourced MLLMs					Closed-sourced MLLMs			
		Qwen-VL	Fuyu†	Blip-2†	InstructBLIP†	LLaVA-1.5	GPT-4V	Gemini†	Claude3 (Sonnet)†	Claude3 (Opus)
Sequence	67.84	18.18	22.42	23.03	23.03	23.64	28.48	23.03	26.06	30.30
Two-row	71.56	16.00	22.67	24.89	24.44	25.78	22.22	24.00	25.78	32.00
Matrix	73.56	23.56	29.78	28.89	28.89	30.67	20.44	26.67	27.56	27.56
Group	58.18	21.48	21.48	21.48	21.48	25.19	17.04	25.93	27.41	24.44
Reassembling	50.00	15.00	15.00	15.00	15.00	15.00	30.00	30.00	20.00	25.00

Table 4: Performance of different models for different task configurations. † notes the result is attributed to inductive bias.

Category	Qwen-VL	Fuyu	Blip2	InstructBLIP	LLaVA-1.5	GPT-4V	Gemini	Claude3 (Sonnet)*	Claude3 (Opus)
Location	40.00	15.38	40.00	35.38	61.54	41.54	41.54	32.31	32.31
Color	37.74	26.42	58.49	58.49	58.49	52.83	54.72	47.17	62.26
Shape	33.33	36.16	45.20	36.72	29.38	45.76	53.67	57.63	45.76
Quantity	36.84	36.84	60.86	56.25	57.57	51.64	39.47	56.58	48.36
Comparison	40.94	40.35	52.05	53.22	57.31	60.82	41.52	42.11	47.95

Table 5: Coarse-grained Perception Performance on Different Categories bias.

formats (sequence and two-row). Table 10 shows the model's few-shot performance on different patterns.

Input	Open-sourced MLLMs					Closed-sourced MLLMs			
	Qwen-VL	Fuyu	Blip-2	InstructBLIP	LLaVA-1.5	GPT-4V	Gemini	Claude3 (Sonnet)	Claude3 (Opus)
AVR	23.16	23.16†	23.16†	23.16†	21.05	21.05	26.32†	27.37†	30.53
AVR+Text	24.21	26.32	26.32	13.68	28.42	65.26	37.89	49.47	55.79

Table 6: Performance of different models after introducing text description in the input. † notes the result is attributed to inductive bias.

Model	zero-shot	one-shot	two-shot
Gemini	25.06	28.7	26.75
GPT-4V	22.34	26.1	28.31
Claude3 (sonnet)	26.49	25.97	26.23
Claude3 (Opus)	28.83	27.79	23.64

Table 7: Few-shot COT Accuracy

Model	one-shot	one-shot (OOD)
Gemini	23.85	24.62
GPT-4V	27.31	23.85
Claude3 (sonnet)	25.00	26.92
Claude3 (Opus)	29.62	25.00

Table 8: Few-shot COT Ablation

Model	two-shot	two-shot (mix)
Gemini	30.48	24.29
GPT-4V	29.52	28.10
Claude3 (sonnet)	28.57	23.81
Claude3 (Opus)	21.90	22.86

Table 9: Few-shot COT Ablation

	one -shot				two -shot			
	Gemini	GPT-4V	Claude3 (Sonnet)	Claude3 (Opus)	Gemini	GPT-4V	Claude3 (Sonnet)	Claude3 (Opus)
Temporal Movement	24.76	27.62	17.14	29.52	33.33	28.57	24.76	18.10
Spatial Relationship	30.00	33.33	24.17	26.67	25.83	22.50	30.83	19.17
Quantities	21.21	27.88	27.88	26.67	25.45	31.52	25.45	27.27
Mathematical	27.92	28.75	25.42	26.67	25.83	27.50	22.92	22.08
2D-Geometry	26.67	25.00	30.83	30.83	26.67	30.00	28.33	31.67
3D-Geometry	25.00	35.00	45.00	30.00	20.00	35.00	40.00	20.00

Table 10: Few-shot performance on different patterns

Experiment	Prompt Example
AVR Question (Reassembling)	*[IMG]* You are given a puzzle. The puzzle consists of a context part on the top and the choices in the bottom. The context part on the top is an unfolded diagram of the 3D shape. The choices part on the bottom contains 4 options (marked by 1, 2, 3, or 4) represents the correct three-dimensional assembly. Which option (either 1, 2, 3, or 4) is the most appropriate answer?
AVR Question (Group)	*[IMG]* You are given a puzzle. The puzzle consists of a context part on the top and the choices at the bottom. The context part on the top is a set of visual patterns arranged in a three-by-one sequence, with the last piece missing. The choices part on the bottom contains four options (marked by 1, 2, 3, or 4). Which option (1, 2, 3, or 4) is the most appropriate answer to fill the missing part?
AVR Question (Matrix)	*[IMG]* You are given a puzzle. The puzzle consists of a context part on the top and the choices at the bottom. The context part on the top is a set of visual patterns arranged in a three by three matrix, with the bottom right piece missing. The choices part on the bottom contains four options (marked by 1, 2, 3, or 4). Which option (1, 2, 3, or 4) is the most appropriate answer to fill the missing part?
AVR Question (Sequence)	*[IMG]* You are given a puzzle. The puzzle consists of a context part on the top and the choices part on the bottom. The context part on the top is a set of visual patterns arranged sequentially, with the last piece missing. The choices part on the bottom contains four options (marked by 1, 2, 3, or 4). Which option (1, 2, 3, or 4) is the most appropriate answer to fill the missing part?
AVR Question (Two-row)	*[IMG]* You are given a puzzle. The puzzle consists of a context part on the top and the choices part on the bottom. The context part on the top is a set of visual patterns arranged in a three by two matrix, with the bottom right piece missing. The choices part on the bottom contains four options (marked by 1, 2, 3, or 4). Which option (1, 2, 3, or 4) is the most appropriate answer to fill the missing part?
Perception Question (fine-grained)	*[IMG]* You are given a puzzle. The puzzle consists of a context part on the top and the choices part on the bottom. The context part on the top is a set of visual grids arranged in a *m by n* sequence, with the last piece missing. The choices part on the bottom contains four choices (marked by 1, 2, 3, or 4). *[question]*
Perception Question (coarse-grained)	*[IMG]* You are given a puzzle. The puzzle consists of a context part on the top and the choices part on the bottom. The context part on the top contains some grids, with the last missing blank grid to be completed. The choices part on the bottom contains a sequence of grids representing the possible choices. How many grids, including the blank grid, are in the context part?
Perception Question (coarse-grained)	*[IMG]* You are given a puzzle. The puzzle consists of a question part on the top and the choices part on the bottom. The question part on the top contains some grids, with the last missing blank grid to be completed. The choices part on the bottom contains a sequence of grids representing the possible choices. How many grids are there in the choices part?
Perception Question (coarse-grained)	*[IMG]* You are given a puzzle. The puzzle consists of a question part on the top and the choices part on the bottom. The question part on the top contains some grids, with the last missing blank grid to be completed. The choices part on the bottom contains a sequence of grids representing the possible choices. How many grids, including the blank grid, are there in the whole puzzle?

Table 11: Prompt examples for our experiments.