Beyond Perception: Evaluating Abstract Visual Reasoning through Multi-Stage Task

Anonymous ACL submission

Abstract

Current Multimodal Large Language Models (MLLMs) excel in general visual reasoning but remain underexplored in Abstract Visual Reasoning (AVR), which demands higher-order reasoning to identify abstract rules beyond simple perception. Existing AVR benchmarks focus on single-step reasoning, emphasizing the end result but neglecting the multi-stage nature of reasoning process. Past studies found MLLMs struggle with these benchmarks, but it doesn't explain how they fail. To address this gap, we introduce MultiStAR, a **Multi-St**age **AVR** benchmark, based on RAVEN, designed to assess reasoning across varying levels of complexity. Additionally, existing metrics like accuracy only focus on the final outcomes while do not account for the correctness of intermediate steps. Therefore, we propose a novel metric, MSEval, which considers the correctness of intermediate steps in addition to the final outcomes. We conduct comprehensive experiments on MultiStAR using 17 representative close-source and open-source MLLMs. The results reveal that while existing MLLMs perform adequately on basic perception tasks, they continue to face challenges in more complex rule detection stages. The dataset and code will be available after acceptance.

1 Introduction

Multimodal Large Language Models (MLLMs) demonstrate proficiency in addressing a wide array of visual-text inquiries and show strong multimodal understanding ability in tasks such as visual question answering (Goyal et al., 2017; Marino et al., 2019; Ding et al., 2023), image captioning (Saito et al., 2023; Vinyals et al., 2015; Jiang et al., 2024a), and visual grounding (He et al., 2024; Deng et al., 2021). These tasks focus on evaluating the models' capability to understand real-world or domain-specific knowledge. However, Abstract Visual Reasoning (AVR) presents a different challenge, focusing on a model's ability to identify and reason

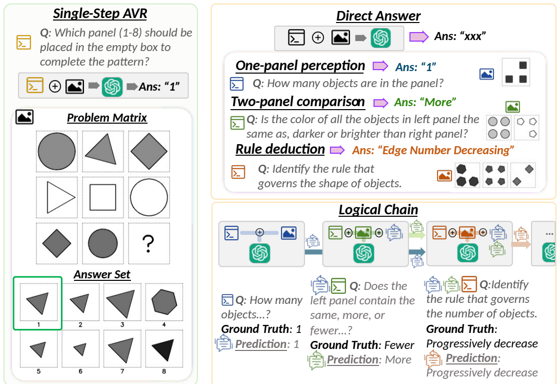


Figure 1: *Left Part*: RAVEN puzzle. The correct answer is 1. *Right Part*: Direct Answer subtask, where questions are independent for each configuration; Logical Chain subtask, where information from previous stages is used to assist in answering the current stage, all questions here focus on the concept of Number.

through abstract patterns, relationships, and rules. A well-known example of AVR tasks is RAVEN (Raven, 2003; Zhang et al., 2019), as shown in the left part of Figure 1. The solver needs to select the correct panel from a answer set to complete a 3x3 problem matrix by deducing the visual rules governing the grid's arrangement. For instance, by analyzing the colors of each panel, one might observe the color remains consistent across each row. Unlike other multimodal tasks in real-world scenarios, AVR focuses on reasoning about arbitrary visual elements, serving as a robust benchmark for evaluating the zero-shot reasoning capabilities of MLLMs in visual contexts (Mańdziuk and Żychowski, 2019; Santoro et al., 2018).

Previous works have consistently shown that AVR tasks pose challenges for MLLMs in zero-shot inference settings. Despite recent advancements like Chain-of-Thought prompting (Ahrabian et al., 2024; Gendron et al., 2024) and the inclusion of oracle captions (Zhang et al., 2024), models continue to perform at near-random levels on these tasks. The AVR datasets used commonly in these evaluations like RAVEN primarily focus on single-step end-to-end reasoning (i.e., giving the models the questions and asking them to derive the final answer), as shown in the left part of Figure 1 (Santoro et al., 2018; Nie et al., 2020; Cao et al., 2024). However, this design deviates from the human reasoning process, which often involves sequential steps: starting with single-panel perception, progressing to panel comparisons, and finally deducing the underlying rules before solving the puzzle. Previous datasets often omit these intermediate stages, posing challenges to effectively evaluate their step-by-step reasoning capabilities and identify where models struggle within the reasoning process. This highlights the need for benchmarks that *assess intermediate perception and reasoning processes*. Additionally, a model that accurately identifies patterns in early steps but fails in the final deduction still demonstrates partial reasoning capability. Rewarding such intermediate success aligns with human evaluation practices. However, existing metrics like accuracy, measure only the performance of the current stage while *disregarding the correctness of intermediate steps.*

To address the limitation of lacking intermediate process evaluation, we introduce MultiStAR, a Multi-Stage Abstract Visual Reasoning dataset, designed to evaluate MLLMs on the intermediate steps in the reasoning process. As shown in Figure 1, the dataset is divided into two sub-tasks, each focusing on different aspects of reasoning. The first sub-task, referred to as **Direct Answer**, evaluates model performance at varying levels of complexity to assess perception and reasoning abilities at each individual step. Using template-based methods, we generate questions based on RAVEN, ranging from basic object recognition to advanced comparison, pattern recognition, and rule inference. This approach ensures comprehensive coverage of reasoning patterns. The second sub-task called **Logical Chain**, emphasizes how models measure and maintain logical correlations across reasoning steps. Using puzzles from the RAVEN as the final question, we decompose the reasoning process into a sequence of subproblems in a bottom-up manner. Each stage in this chain links the current reasoning task to its dependent subproblems, requiring the model to combine current information with outputs from previous stages. To assess the correctness of intermediate steps, we introduce a novel metric, MSEval, which provides a more fine-grained assessment of the model's reasoning process for the logical chain task. MSEval uses the correct answer probabilities at each stage to compute the joint probability across the reasoning process. This approach considers both the correctness of the current stage and all dependent intermediate steps.

In summary, our contributions are: 1) We introduce the MultiStAR benchmark, designed to evaluate models across different stages of reasoning through two subtasks, allowing for a more granular analysis of their performance throughout the reasoning process. 2) We present a novel metric that incorporates the correctness of the current stage as well as the accuracy of its dependent intermediate steps. 3) We perform extensive experiments on a wide range of state-of-the-art MLLMs, providing insights into their strengths, weaknesses, and future improvement directions on AVR tasks.

2 Related Work

Visual reasoning benchmarks have evolved to assess the capacity of AI models on various tasks including compositional (Johnson et al., 2017), commonsense (Gao et al., 2022; Li et al., 2024), scientific (Hiippala et al., 2021; Saikh et al., 2022; Yue et al., 2024), and abstract visual reasoning. Both commonsense and scientific reasoning tasks require real-world knowledge and a prior understanding of specific domains. Abstract Visual Reasoning (AVR) benchmarks the main focus of this work, primarily involving classification tasks, where models select an answer from a fixed set of choices based on abstract patterns and rules (Mańdziuk and Żychowski, 2019; Zhang et al., 2019; Santoro et al., 2018; Nie et al., 2020). A few other AVR benchmarks address generative tasks, where models are tasked with recreating elements that fit within a given visual sequence, introducing additional complexity by evaluating a model's creative reasoning capabilities (Chollet, 2019; Moskvichev et al., 2023). The most similar benchmark to ours is MARVEL (Jiang et al., 2024b), which targets AVR tasks and extends reasoning diversity with six core patterns across geometric and abstract shapes. It also includes basic perception questions to assess visual comprehension. However, MARVEL is still limited in its capacity to analyze intermediate reasoning steps. For the key statistics and features comparison of major multimodal reasoning datasets alongside our proposed MultiStAR benchmark, please see Appendix A.1.

3 Multi-stage Evaluation Benchmark

3.1 Task Definition and Configuration

Our dataset comprises two sub-tasks, Direct Answer and Logical Chain, both derived from RAVEN

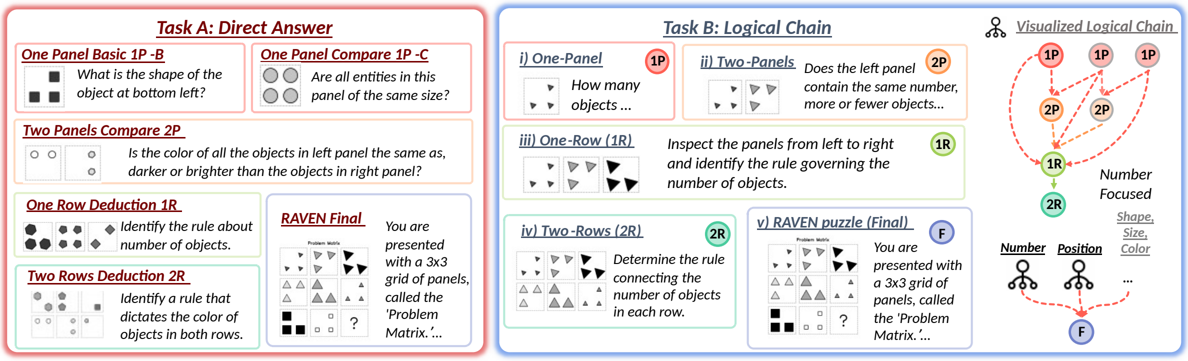


Figure 2: *Left Part:* Direct Answer subtask, showcasing six configurations along with their corresponding examples. *Right Part:* Logical Chain task, presenting a partial view of the logical chain (See full chain and the chain designing rationale in Appendix A.3). Examples are provided for **one specific path** in the chain.

but with distinct reasoning patterns and focuses.

3.1.1 Direct Answer

To uncovering where the MLLMs likely to succeed or struggle *in the individual stages*, this sub-task explores AVR across multiple levels, which is divided into six configurations, shown in Figure 2:

*a) One Panel Basic Perception (**1P-B**):* The puzzle image consists of a single panel $I = \mathbf{p}$, focusing on basic perception questions, such as determining the number of objects, the shape, or the position of a single object, without requiring any comparison.

*b) One Panel Comparison (**1P-C**):* The puzzle image remains a single panel, but questions require intra-panel attribute comparisons.

*c) Two Panels Comparison (**2P**):* The puzzle image consists of two panels, $I = (\mathbf{p}_1, \mathbf{p}_2)$, requiring cross-panel comparisons.

*d) One Row Rule Deduction (**1R**):* The puzzle image is a single row of three panels, $I = (\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3)$, and the task involves identifying a rule that governs the sequence.

*e) Two Row Rule Deduction (**2R**):* The puzzle image contains the first two rows, each with three panels, denoted as $I = (\{\mathbf{p}_{1,1}, \mathbf{p}_{1,2}, \mathbf{p}_{1,3}\}, \{\mathbf{p}_{2,1}, \mathbf{p}_{2,2}, \mathbf{p}_{2,3}\})$. The task is to find a rule that applies to both rows.

*f) RAVEN puzzle (**Final**):* The original puzzle from RAVEN dataset.

Formally, given an puzzle image I (which consist of one or more panels $\mathbf{p}$) and a question q, the task is to select an answer a from a set of k multiple-choice options:

$$a^* = \arg\max_{a \in \mathcal{A}} P(a \mid I, q) \tag{1}$$

where $\mathcal{A} = \{a_1, \ldots, a_k\}$ is the answer set.

3.1.2 Logical Chain

To measure the *sequential steps* of the reasoning process required to reach the final answer, rather than evaluating stages in isolation, the second Logical Chain task extends reasoning across multiple subproblems, introducing dependencies between stages to form a coherent logical chain. As illustrated in the right part of Figure 2, each node represents a stage question, and edges are connected if the previous information is necessary to answer the current stage.

This task consists of five stages, similar to Direct Answer subtask: ***1P*** (Merged 1P-B and 1P-C), ***2P***, ***1R***, ***2R*** and ***Final***. Specifically, each node t involves predicting an answer a_t based on the current question q_t, the current image I_t, and information from prior stages $\mathbf{H}_t$, the task is defined as:

$$a_t^* = \arg\max_{a_t \in \mathcal{A}_t} P(a_t \mid I_t, q_t, \mathbf{H}_t) \tag{2}$$

$$\mathbf{H}_t = \{\text{Re-Format}(q_j, a_j) \mid j \in \mathcal{D}_t\} \tag{3}$$

where $\mathbf{H}_t$ represents the set of prior information, as determined by the pre-defined logical chain $\mathcal{D}_t$, specifying one or more nodes that current node t depends on. As the images referenced by prior questions change across different stages, we use a rule-based program to reformat each dependent question q_j and the generated answer a_j, appending this prior information before the current question to construct the input for the current node. Details of this program are provided in Appendix A.4.2.

3.2 Dataset Creation

Data Sources: Our MultiStAR dataset is derived from the RAVEN dataset, which its associated XML files provide objects details and ground-truth logical rules for generating each puzzle.

Define Templates: We pre-define question templates for all six configurations, each template including a question format, constraints, an answer space, and a corresponding function sequence, as illustrated in Figure 3. Overall, we created 25 distinct templates, details are shown in Appendix A.2.

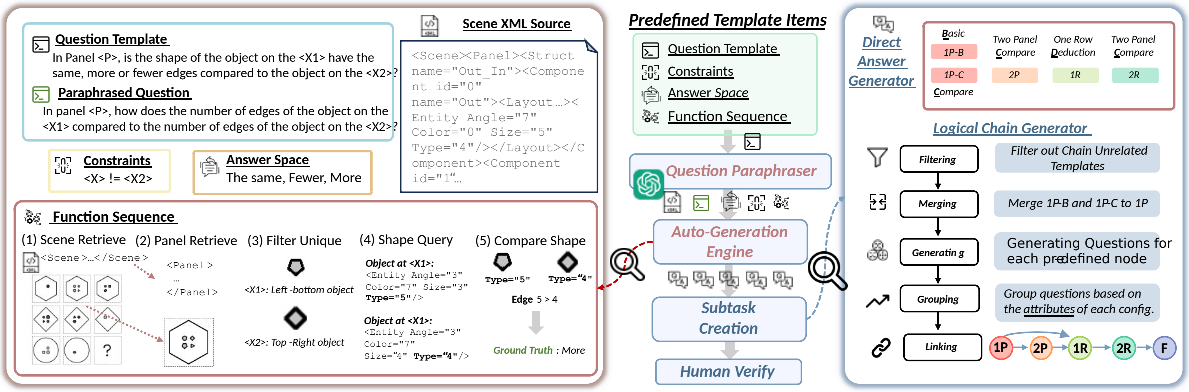


Figure 3: Our MultiStAR dataset generation pipeline.

Question Generation: By leveraging the puzzle information and the pre-defined templates, we implement an automated template-based generation process to efficiently produce large-scale question-answer pairs. Firstly, to enrich question formats and linguistic diversity, we employ GPT-4o (OpenAI, 2024) to rewrite the templates. Then, following a methodology similar to the CLEVR dataset (Johnson et al., 2017), we design functional programs that execute a sequence of functions. For instance, as shown in Figure 3, the program "Scene Retrieve → Panel Retrieve → Filter Unique → Shape Query → Compare Shape" identifies the puzzle matrix, retrieves the relevant panel <P>, locates objects at positions <X1> and <X2>, queries their shapes, and compares them to determine the ground-truth answer. And lastly, the multiple-choice options are sampled from the answer space.

Subtasks Creation: To create the Direct Answer subtask, we first sample XML files from RAVEN, and for each XML file, we generate one question for each *template*. During question formation, placeholders (e.g., <X1>, <X2>) are replaced with randomly selected any possible values consistent with the value ranges and constraints. Next, we create the Logical Chain sub-task by first filtering out those templates that do not contribute useful information for building the logical chain (i.e. the question does not provide necessary input for its child nodes). To simplify chain construction, the first two one-panel configurations, 1P-B and 1P-C, are combined into a single stage representing one-panel information. During question generation, one question is created for each *node*, with placeholders such as panel <P> replaced by the values corresponding to the current node's position in the chain. For instance, if there are three 1P nodes in the chain, they correspond to panels 1, 2, and 3, respectively. Questions are then grouped by attributes such as number and position, aligning with how the chain is constructed. Finally, we assign the previous nodes for each question to establish the edges between nodes. Detailed analysis of our dataset MultiStAR, such as the question length and function distribution, please see Appendix A.1. In total, we create 21.7K questions for Direct Answer subtask, and 3.92K for Logical Chain subtask.

Human Verification: To evaluate the quality of automatically generated question-answer pairs, we also conduct human study based on three aspects, *Correctness*, *Clarity* and *Content Validity*. The results show our dataset performing well across all aspects, see Appendix A.5.1 for details.

3.3 Evaluation Metrics

We use accuracy for the Direct Answer subtask, as it directly aligns with the task of selecting the correct answer from multiple choices. However, for the **Logical Chain** subtask, accuracy alone does not consider intermediate reasoning steps, focusing only on the end result. To address this limitation and better align with the step-by-step reasoning process, we introduce a new metric, MSEval. As the example illustrated in Figure 4, the score for the 1R node is designed to aggregate from all its related nodes, which include three 1P nodes, two 2P nodes, and the 1R node itself. This aggregation captures the interconnectivity between nodes in the logical chain. And to reflect their contribution to the reasoning process, it assigns a weight to each of these nodes based on their importance. Specifically:

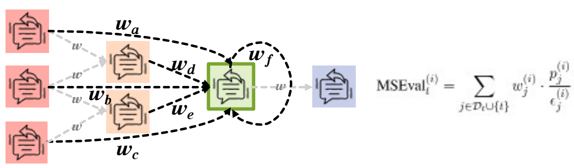


Figure 4: This example is the MSEval score calculation for the 1R node, depends on 1P, 2P and 1R itself. The corresponding weights are denoted as w_a through w_f.

Aggregated Outcomes: To aggregated the intermediate steps outcomes across the chain, MSEval chooses to compute a joint probability of the current node and all dependent nodes as the product of their conditional probabilities. To prevent disregarding the model's performance when it is incorrect, each conditional probability is derived by measuring the probability assigned to the ground truth answer. Using logits from the model's final layer for the answer choices (e.g., A, B, C, D), the correct answer logit $p_j^{(i)}$ is transformed into a probability via the softmax function. This is defined as:

$$\begin{aligned}\text{Joint Prob}_t^{(i)} &= \text{Norm}(P(a_t^{(i)} = a_t^{(i)*}, \Phi_t^{(i)} \mid \xi_t^{(i)})) \\ &= \prod_{j \in \mathcal{D}_t \cup \{t\}} \exp(\frac{p_j^{(i)}}{\epsilon_j^{(i)}}) \end{aligned} \quad (4)$$

$$p_j^{(i)} = P(a_j^{(i)} = a_j^{(i)*} \mid \mathbf{H}_j^{(i)}) = \frac{\exp(z_{a_j^{(i)*}})}{\sum_{k \in \mathcal{A}_j^{(i)}} \exp(z_k)} \quad (5)$$

where $z_{a_j^{(i)*}}$ is the logit for the correct answer $a_j^{(i)*}$, and $\mathcal{A}_j^{(i)}$ is the set of all possible answers, $\epsilon_j^{(i)} = \frac{1}{|\mathcal{A}_j^{(i)}|}$ represents the random rate. The term $\exp(\cdot)$ normalizes the probability to account for varying numbers of choices for all nodes. $\Phi_t^{(i)} = \{a_j^{(i)} = a_j^{(i)*} \mid j \in \mathcal{D}_t\}$ represents the correct answer probability for all dependent nodes, $\xi_t^{(i)} = \{\mathbf{H}_j^{(i)} \mid j \in \mathcal{D}_t \cup \{t\}\}$ represents a set of prior information for each node.

Weighted Importance: The joint probability does not account for the relative importance of each node in the chain. To address this, we introduce a weight for each node based on its influence on the current node, as measured by conditional mutual information (CMI). CMI is obtained by altering the set of all possible answers $\mathcal{A}_j^{(i)}$ at node j while keeping the outputs of all other nodes ($\mathcal{D}_t^{(i)} \setminus \{j\}$) fixed. We then observe how the model's outputs $\mathcal{A}_{j\to t}^{(i)}$ at the current node t change. If $\mathcal{A}_{j\to t}^{(i)}$ changes significantly, the CMI is higher, resulting in a higher weight. As raw CMI values may vary in scale, we take a normalization of the conditional mutual information (NCMI). This process is defined as:

$$\begin{aligned}\text{CMI}(i,j,t) &= \text{CMI}(\mathcal{A}_j^{(i)}; \mathcal{A}_{j\to t}^{(i)} \mid \mathcal{D}_t^{(i)} \setminus \{j\}) \\ &= H(\mathcal{A}_{j\to t}^{(i)} \mid \mathcal{D}_t^{(i)} \setminus \{j\}) \\ &\quad + H(\mathcal{A}_j^{(i)} \mid \mathcal{D}_t^{(i)} \setminus \{j\}) \\ &\quad - H(\mathcal{A}_{j\to t}^{(i)}, \mathcal{A}_j^{(i)} \mid \mathcal{D}_t^{(i)} \setminus \{j\}) \end{aligned} \quad (6)$$

$$\text{NCMI}(i,j,t) = \frac{\exp(\text{CMI}(i,j,t))}{\sum_{k \in \mathcal{D}_t \cup \{t\}} \exp(\text{CMI}(i,k,t)))} \quad (7)$$

where $H(\cdot)$ denotes entropy. Note that for current node t, we have $\mathcal{A}_t^{(i)} = \mathcal{A}_{t\to t}^{(i)}$, so that current node would always have the highest impact to itself.

We apply the NCMI to each node's conditional probability to compute its weighted contribution to the reasoning process. To simplify the formulation, we apply a log to the expression. The final **MSEval** score for stage t, instance i is computed as:

$$\begin{aligned}\text{MSEval}_t^{(i)} &= \log \prod_{j \in \mathcal{D}_t \cup \{t\}} (\exp(\frac{p_j^{(i)}}{\epsilon_j^{(i)}}))^{\text{NCMI}(i,j,t)} \\ &= \sum_{j \in \mathcal{D}_t \cup \{t\}} \text{NCMI}(i,j,t) \cdot \frac{p_j^{(i)}}{\epsilon_j^{(i)}} \end{aligned} \quad (8)$$

$$\text{MSEval}_t^{(i)} = \sum_{j \in \mathcal{D}_t \cup \{t\}} w_j^{(i)} \cdot \frac{p_j^{(i)}}{\epsilon_j^{(i)}} \quad (9)$$

As MSEval relies on access to the logits of the model's final layer, which can only be applied to **open-source** models. More details about MSEval, such as algorithm Pseudo Code and computational cost, are shown in Appendix A.7.

4 Results

4.1 Experiment Setup

In the Direct Answer subtask, we evaluate 17 representative MLLMs in zero-shot setting, including both open-source and close-source models. For open-source models, we have pre-trained (without instruction-tuned) and instruction-tuned models. Additionally, we consider a range of model sizes to ensure a comprehensive evaluation. The model settings and the input prompt details are shown in Appendix A.4.1.

In the Logical Chain subtask, we evaluate six models which performed well in the Direct Answer subtask in zero-shot setting. To compare how much benefit the models gain from prior information, we also evaluate the models without providing prior information ($\mathbf{H}_t = \emptyset$), serving as a baseline for comparison. In this case, as we only consider the current stage, the MSEval score simplifies to $\text{MSEval}_t^{(i)} = \frac{p_t^{(i)}}{|\mathcal{A}_t^{(i)}|}$. The model settings and the input prompt details are shown in Appendix A.4.2.

To establish the human performance on the subproblems within MultiStAR, we conduct a human study on a crowd-sourcing platform in which participants solved a 10% subset of the benchmark. We do not evaluate human performance on the original RAVEN puzzles, but instead use the result reported in Zhang et al. (2019). Please see Appendix A.5.2 for more details about the human study.

4.2 Result Analysis

Direct Answer: Table 1 compares the performance of various MLLMs on the Direct Answer task. The two close-source models, GPT-4o and Gemini-1.5-pro, outperform others in 1P-B, 1R and 2R stages. GPT-4o achieves an impressive 88.07% accuracy for basic object-oriented questions within a single panel, highlighting its strong capability to recognize simple visual patterns. Gemini achieves the best results in rule deduction tasks for both one-row and two-row configurations, indicating its superior ability to process more complex visual inputs and perform logical reasoning effectively. Among open-source models, pre-trained models generally perform worse than instruction-tuned ones, emphasizing that instruction tuning is an effective approach for addressing question-answering and reasoning tasks. Qwen2-VL-Instruct-72B achieves the best performance on cross-panel comparison tasks, showcasing its strength in identifying relationships between objects across panels. For the final stage, models like Qwen2-VL-Instruct and Idefics2-8B achieve significantly higher scores. We suspect that these models may have encountered RAVEN questions during their training process.

Interestingly, as the questions become increasingly complex and require deeper reasoning, a noticeable decline in performance is observed across all models, gradually approaching the random baseline. While models demonstrate strong performance on basic perception tasks, they struggle significantly with deeper reasoning challenges. In contrast, human performance remains stable at above 60% with increasing complexity. This highlights the substantial gap between model and human performance, emphasizing the limitations of current MLLMs in understanding and reasoning at a level comparable to humans.

Another finding is that the performance generally improves with larger model sizes (See Appendix A.8.1 for detailed visual analysis.), mainly due to differences in the size of their language encoders. This indicates that a robust language encoder significantly influences overall performance.

Logical Chain: Table 2 shows the performance of MLLMs on each stage of the Logical Chain task. Results are reported as accuracy and MSEval scores. Among all models, Gemini-1.5-pro achieves the best performance on the first four stages, demonstrating its superior ability to reason through multi-stage dependencies. Among the open-source models, Qwen2-VL-72B outperforms others in both accuracy and MSEval, suggesting that our MSEval metric generally aligns with accuracy. When comparing results with and without prior, all models show better performance when prior is available in both metrics, highlighting their capacity to benefit from step-by-step reasoning, even when some generated previous answers might be incorrect. For visual representation of the percentage increase, refer to Appendix A.6.

Interestingly, for the final stage involving the RAVEN puzzle, prior information appears to provide limited utility for most models, with accuracy close to random except for those that might touch on RAVEN tasks. This aligns with previous findings that chain-of-thought reasoning models struggle to solve RAVEN puzzles (Ahrabian et al., 2024; Gendron et al., 2024). However, MSEval scores tell a different story. Models with prior information, particularly NVLM-D-72B, show significant improvements in MSEval (close to 100%), despite low final stage accuracy. Since MSEval evaluates correctness across intermediate and current stages, it reveals that models, despite failing the final stage, often solve intermediate steps with higher confidence. Another finding is the MSEval score for Qwen2-VL-72B declines when provided with prior information, which is not consistent with accuracy. This indicates its weakness in addressing intermediate stages that were likely not part

	1P-B	1P-C	2P	1R	2R	Final
Close-Source Models						
GPT-4o (2024)	**88.1**	72.7	54.0	40.0	31.6	12.1
Gemini-1.5-pro (2023)	83.2	75.0	50.0	**46.9**	**37.8**	11.6
Pre-trained Open-Source Models						
Qwen-VL-7B (2023)	17.5	24.7	22.5	15.8	12.2	12.3
Idefics2-8B (2024)	17.2	33.0	27.3	19.8	21.4	12.3
xGen-MM-4B (2024)	40.2	31.8	12.5	24.1	23.9	3.4
Instruction-Tuned open-source Models						
Instructblip-7B (2023)	27.5	37.1	27.7	14.0	13.5	11.6
Instructblip-13B (2023)	29.4	39.0	26.9	25.0	23.0	14.3
LLaVA-v1.5-7B (2024)	47.8	50.2	32.9	27.1	25.7	13.3
LLaVA-v1.5-13B (2024)	59.6	47.6	15.9	26.7	26.9	11.3
Idefics2-8B (2024)	85.1	65.5	42.0	37.0	36.8	29.9*
xGen-MM-4B (2024)	81.2	47.9	21.9	24.0	25.6	2.4
Qwen2-VL-2B (2024)	40.4	42.7	22.8	13.7	11.4	9.9*
Qwen2-VL-7B (2024)	64.8	56.6	47.9	31.0	33.2	24.3*
Qwen2-VL-72B (2024)	86.9	**77.8**	**60.2**	45.5	21.9	**63.7***
NVLM-D-72B (2024)	80.5	67.1	45.3	39.1	31.3	12.7
Intern-VL2-2B (2024)	54.0	48.7	27.3	26.9	23.9	10.1
Intern-VL2-8B (2024)	63.0	54.5	34.2	23.2	23.4	14.6
Random	39.9	23.0	26.2	25.0	25.0	12.5
Human	98.5	88.9	69.1	62.1	63.3	84.4‡

Table 1: The answer accuracy of MLLMs for the Direct Answer subtask. The best results are highlighted in bold. *The model may have included the RAVEN dataset in training. ‡Result reported in (Zhang et al., 2019).

	Metric	Prior	1P	2P	1R	2R	Final
GPT-4o	Acc	w/o	73.8	39.1	34.7	28.9	15.7
		w	73.8	43.9	41.8	50.6	10.0
Gemini	Acc	w/o	75.5	61.6	49.6	44.6	5.7
		w	**75.5**	**64.4**	**52.6**	**57.1**	18.6
Idefics2 (8B)	Acc	w/o	57.8	39.6	34.4	35.1	25.7*
		w	57.8	37.8	36.6	42.4	25.7*
	MSEval	w/o	2.02	1.29	1.29	1.27	1.24*
		w	2.02	1.48	1.51	1.51	1.44*
Qwen2-VL (72B)	Acc	w/o	74.1	56.0	45.1	42.9	64.3*
		w	74.1	57.8	47.3	54.2	**65.7***
	MSEval	w/o	2.54	1.95	1.79	1.70	<u>5.14</u>*
		w	<u>2.54</u>	<u>2.13</u>	<u>2.12</u>	<u>2.10</u>	3.31*
Intern-VL2 (8B)	Acc	w/o	54.4	35.9	21.6	21.6	18.6
		w	54.4	41.9	31.6	33.5	17.1
	MSEval	w/o	1.75	1.09	0.90	0.91	1.18
		w	1.75	1.38	1.30	1.18	1.26
NVLM-D (72B)	Acc	w/o	66.1	42.4	37.8	23.5	8.6
		w	66.1	45.2	39.1	43.3	7.1
	MSEval	w/o	2.25	1.20	1.28	1.02	0.76
		w	2.25	1.65	1.69	1.62	1.41
Random	Acc	-	31.1	31.7	25.0	25.0	12.5
	MSEval	-	1.00	1.00	1.00	1.00	1.00

Table 2: The Accuracy (Acc) and MSEval scores for the Logical Chain task. *The model may have included the RAVEN dataset in training. The highest accuracy are highlighted in **bold**. The highest MSEval are highlighted in <u>underline</u>. w/o: without prior, w: with prior

of its pre-training. Despite errors in earlier stages, the model still performs well on the final question, suggesting it relies on memorizing patterns from the final stage rather than demonstrating a strong understanding of the logical reasoning behind the task. This highlights a critical limitation in current MLLMs, **while they may achieve impressive results in isolated cases, their ability to generalize and reason through multi-stage logical dependencies remains inadequate**. To further verify the MSEval's effectiveness, we also conduct qualitative analysis in section 6.

5 Discussion

What insights can be drawn from each attribute's performance? From Figure 5, the "number" attribute is the easiest to recognize in higher level configurations (2P, 1R, 2R), while "position" is the most easily identified in low-level, single-panel settings. Some attributes achieve accuracy above 90%, indicating that the models exhibit strong counting and spatial reasoning capabilities. However, they struggle with attributes like "color" and "size", particularly in high-level configurations, suggesting that the models may not be adequately designed or trained to focus on these aspects.

Given the ground truth for intermediate steps, how does it influence the final results? Table 3 highlights the ground truth priors generally demonstrate a positive impact. For example, the 1R stage

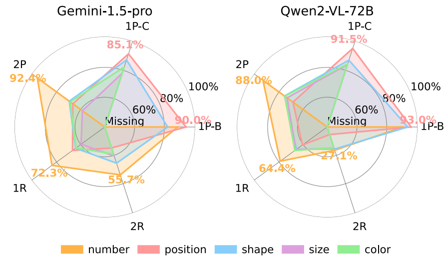


Figure 5: Breakdown analysis of five attributes for Gemeni-1.5-pro and Qwen2-VL-72B on the Direct Answer task. Refer to the Appendix A.8.2 for more models.

benefits significantly from the insights about each panel and intra-panel comparisons. The 2R stage also sees substantial gains, as it mainly relies on double-checking information from the 1R stage without requiring additional changes in most cases. However, the final stage experiences a negative impact despite the inclusion of correct rules. This may be attributed to the complexity of the visual input, which contains numerous objects, making it challenging for the model to effectively apply the given rules. And for the Qwen2-VL-72B model, its tendency to memorize patterns might turn these ground truths into noise.

	Prior Info	1P	2P	1R	2R	Final
GPT-4o	w/o	73.8	39.1	34.7	28.9	15.7
	GT	73.8	49.2	66.0	93.8	14.3
Gemini	w/o	75.5	61.6	49.6	44.6	5.7
	GT	**75.5**	**62.4**	68.9	79.5	7.1
Qwen2-VL (72B)	w/o	74.1	56.0	45.1	42.9	64.3*
	GT	74.1	61.9	**70.7**	92.2	**55.7***
NVLM-D (72B)	w/o	66.1	42.4	37.8	23.5	8.6
	GT	66.1	51.9	66.4	**96.0**	7.14

Table 3: The accuracy with incorporating ground truth information at each stage for Logical Chain task.

How much do previous stages influence the current stage? A key step in our MSEval metrics is measuring the relative importance of intermediate dependent stages to the current stage using NCMI. This allows us to assess how each step in the chain depends on prior stages, *helping verify whether previous information is useful and if the designed chain is logically sound*. As shown in Table 4, prior information often has significant weight on the current stage, except for the position attribute at "1P to 2P". This suggests that querying object position in a single panel has little impact on determining if positions are the same across panels.

How do variations in handling long prompts affect model outcomes? Injecting prior information into prompts significantly increases their length (see Appendix A.1 for details), making it more challenging for models to focus on critical details. To address this issue, we proposed two methods: (1) adding HTML tags to structure the prompt

	Attributes	1P to 2P	(1P,2P) to 1R	1R to 2R	2R to F
Qwen2-VL (72B)	Number	0.40	0.54	0.33	0.42
	Position	0.21	0.48	0.39	
	Shape	0.37	0.51	0.27	
	Color	0.36	0.53	0.33	
	Size	0.37	0.51	0.32	
NVLM-D (72B)	Number	0.40	0.56	0.37	0.63
	Position	0.21	0.50	0.42	
	Shape	0.40	0.52	0.35	
	Color	0.38	0.54	0.38	
	Size	0.39	0.54	0.38	

Table 4: The summation of NCMI weight assigned to all dependent stages, grouped by each attribute, averaged across instances.

by separating prior information, background, and questions, enabling the model to clearly distinguish each part, and (2) formatting the prompt as a PDF document with distinct sections and titles. Table 5 demonstrates that HTML tagging provides notable benefits, particularly Qwen2-VL and GPT-4o, while the document-based approach proves less effective, especially for high-level stages. However, for other open-source models, they yields no improvements (see Appendix A.8.3 for further details and examples of the conversion methods).

	Prior	1P	2P	1R	2R	Final
GPT-4o	Vanilla	73.8	43.9	41.8	50.6	10.0
	Struct.	**82.2**	64.4	47.8	50.9	8.6
	Doc.	80.8	44.8	31.1	24.9	10.0
Gemini	Vanilla	75.5	64.4	52.6	57.1	18.6
	Struct.	70.6	66.4	52.9	**57.8**	17.1
	Doc.	69.6	51.0	36.7	33.1	14.3
Qwen2	Vanilla	74.1	57.8	47.3	54.2	65.7*
	Struct.	77.2	**67.7**	**55.1**	53.6	61.4
	Doc.	76.5	63.1	50.2	46.6	24.3

Table 5: The accuracy of three prompting techniques for prior information for Logical Chain task. *Vanilla*: Pure Text, *Struct.*: Structure (HTML), *Doc.*: Document.

6 Qualitative Analysis

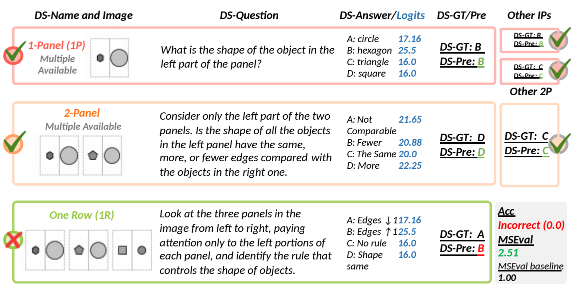


Figure 6: The top two rows are dependent stages (All Correct), the bottom row is current stage (Incorrect).

To highlight the advantages of our MSEval metric over traditional accuracy, we provide several concrete examples across different scenarios. Figure 6 shows a case where the final answer to the current question is incorrect, resulting in an accuracy score of 0.0. However, the model demonstrates strong performance in intermediate steps, correctly solving the one-panel and two-panel comparisons with high confidence. Additionally, the logits for the correct answer "A" are only slightly lower than the highest logits. By considering these factors, the MSEval metric assigns a reasonable score, reflecting the model's partial success. Further examples, including cases where the current question is correct but intermediate steps are incorrect, are provided in Appendix A.10.

7 Error Analysis

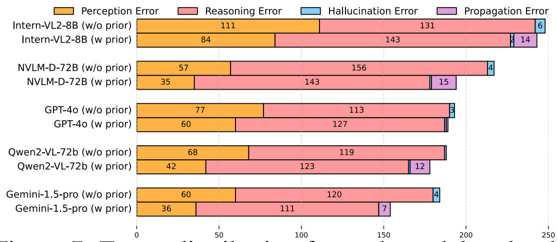


Figure 7: Errors distribution for each model under the settings of without and with prior.

To further investigate model performance, we conduct an error analysis for the Logical Chain subtask. Models are asked to generate with explanations alongside answers, and we manually reviewed all outputs. Errors are classified into four types: (1) Perception Error, misinterpretation of visual inputs such as object numbers or shapes, (2) Reasoning Error, incorrect logic based on correctly perceived inputs, (3) Propagation Error, failure to interpret or detect inaccuracies in prior information, and (4) Hallucination Error, generation of incomplete sentences or unrelated information. Figure 7 reveals that reasoning errors are the most common, followed by perception errors, with hallucination and propagation errors being rare. Gemini exhibits the lowest perception error rate, while GPT-4o shows the lowest reasoning error rate. Notably, injecting prior information significantly reduces perception errors, demonstrating that prior knowledge enhances models' understanding of visual inputs, but it does not help with reasoning error. See Appendix A.9 for concrete examples of each error.

8 Conclusion

In this work, we propose the MultiStAR benchmark. While current models perform well on basic perception tasks, they face significant challenges with deeper reasoning stages. Our findings also indicate that models have the potential to benefit from step-by-step reasoning. However, despite extensive training yielding impressive results in isolated scenarios, their ability to handle logical dependencies remains limited. We introduce a metric MSEval, which can be applied to a variety of reasoning tasks beyond visual reasoning, including domains such as mathematics and science, where multi-step logic is critical, provided there are clearly defined chains.

9 Limitation

The automatic generation methods we use are restricted to datasets with clearly defined object attributes, such as the XML files provided by RAVEN. This limits our expansion to RAVEN dataset, as most datasets lack such metadata. Expanding these methods to other datasets will require machine learning approaches, such as automatic object boundary detection, which could eliminate the need for metadata files.

The logical chain design in our dataset is not perfect. In some cases, prior information is insufficient for the current stage, such as instances in the one-row rule deduction stage where the rule might involve "Three Different Numbers", in this case, we also need the information about the second row. To make the chain construction more easily, currently, we design chains at the "Corpus-Level," meaning they are fixed across all instances. Future work could explore automatic "Instance-Level" chain construction methods, enabling models to dynamically generate chains based on patterns within individual examples.

As the results show (especially in the final stage), current models still lack the ability to navigate multi-stage logical dependencies effectively, even when trained on the data. We do not address this issue in the current work. Future research could explore optimization methods that focus on improving intermediate reasoning steps, rather than just the final outcome, to enhance models' multi-step reasoning capabilities.

Currently, we simply inject the prior information using pure text, HTML tags and treat it as part of the document. However, there are numerous other possibilities, such as using a chat-based interface or multi-turn iterative inputs, to help the model better understand and utilize the prior information.

References

Kian Ahrabian, Zhivar Sourati, Kexuan Sun, Jiarui Zhang, Yifan Jiang, Fred Morstatter, and Jay Pujara. 2024. The curious case of nonverbal abstract reasoning with multi-modal large language models. In *Proceedings of Thirty Seventh Conference on Learning Theory*.

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

Xu Cao, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Jianguo Cao, and James M Rehg. 2024. What is the visual cognition gap between humans and multimodal llms? *arXiv preprint arXiv:2406.10424*.

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

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

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779.

Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. Vqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 585–601. Springer.

Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2022. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5561–5578.

Gael Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large language models are not strong abstract reasoners yet. In *ICLR 2024 Workshop: How Far Are We From AGI*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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*, pages 6904–6913.

Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C Berg, and Vicente Ordonez. 2024. Improved visual grounding through self-consistent explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13095–13105.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. 2021. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688.

Yanbei Jiang, Krista A Ehinger, and Jey Han Lau. 2024a. Kale: An artwork image captioning system augmented with heterogeneous graph. *arXiv preprint arXiv:2409.10921*.

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

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

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

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

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

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

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

OpenAI. 2024. https://openai.com/index/hello-gpt-4o/.

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

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Adam Santoro, Felix Hill, David Barrett, Ari Morcos, and Timothy Lillicrap. 2018. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pages 4477–4486.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

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

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5317–5327.

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

A Appendix

A.1 Dataset Analysis

Table 6 illustrates the statistics of our dataset compared to other multimodal visual reasoning VQA datasets. Figure 8a compares question lengths across different reasoning VQA datasets. Our dataset stands out with a roughly even distribution of question lengths, unlike other datasets that predominantly focus on shorter questions, which makes our dataset more challenging for MLLMs. Figure 8b illustrates the proportion of functional programs used in the dataset, showing a wide variety of functions, with query_rule being slightly more frequent. Figure 8c highlights the number of multiple-choice options for each configuration, where differences in the number of choices arise due to constraints in the answer space for some configurations. Figure 8d presents the input prompt length for each stage in the logical chain task, comparing settings with and without prior information. Incorporating prior information from earlier stages significantly increases the maximum prompt length to 261.2 tokens, posing a challenge for MLLMs to parse effectively.

A.2 Generation Template

Table 7 outlines the question templates used for the Direct Answer Task. Notably, it is impractical to ask questions about the color or size of a single object, as these attributes are represented by numerical values (e.g., color as 255 or size as 8), which models cannot interpret meaningfully. Therefore, questions about size and color are excluded from the basic one-panel tasks. Instead, comparative questions such as "darker" or "smaller" are included in the one-panel comparison tasks. Two constraints are applied during template creation: Not_Equal(P, P2) ensures that Panels P and P2 are different, and Same_Row(P, P2) ensures that Panels P and P2 belong to the same row. For answer spaces exceeding four options, a sampling method is used to limit the choices to a maximum of four. Table 8 lists all possible values that each placeholder can take, as well as the complete set of rules for each attribute in the rule deduction configuration. After the placeholder values are assigned, references to "Panel <P>" are replaced with "this panel" to enhance clarity and readability.

A.3 The Full Logical Chain

Figure 9 presents the full view of our pre-defined logical chain, while Table 9 provides corresponding question examples for each node in the chain. The chain's ultimate objective is to solve the original RAVEN puzzle, where each row contains rules for attributes such as number, position, shape, size, and color. Intuitively, we connect each attribute's rule deduction phase to the final phase, operating under the assumption that knowing all hidden rules of the RAVEN puzzle provides sufficient information to solve it.

To determine the rules, we expand the reasoning scope from one row to two rows. For one-row rule deduction, we link single-panel perception and two-panel comparison to ensure that with panel-level details and inter-panel comparisons, the rule can be identified. This logical chain is crafted to mimic human problem-solving behavior: focusing first on single-panel perception, followed by panel comparisons, then deducing the first-row rule and validating it with the second row.

However, this handcrafted chain design has limitations. First, it may not align with the model's actual reasoning process, which can cause discrepancies in performance. Additionally, to simplify chain construction, we designed it at the "Corpus-Level," meaning it remains fixed across all instances. This approach sometimes results in insufficient prior information for certain stages. For example, in one-row rule deduction, a rule like "the number of objects distributes three distinct values across panels, rotating through each possible permutation" may require second-row information to resolve. These limitations highlight the need for more flexible and instance-specific logical chain designs in future work.

A.4 Input Prompt And Model Settings

A.4.1 Direct Answer

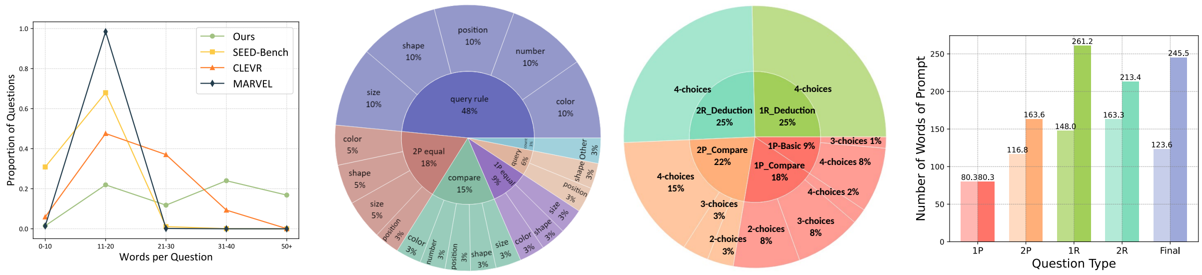
Model Settings: All MLLMs are tested under their default settings under the environment of Huggingface [1], the *transformer* package version in python is 4.39.2 for NVLM-D-72B model and 4.46.0 for all others.

Prompt Details: The RAVEN dataset includes various puzzle settings, such as Left-Right, Up-Down, and In-Out, where rules are applied separately to distinct parts of the panels (Figure 10).

[1] https://huggingface.co/

Dataset	Num. of Images	Num. of QA pairs	Reasoning Domain (Task Focus)	Question Generation	Answer Type	Functional Program	Multi-Step Structure
CLEVR (Johnson et al., 2017)	100K	1M	Compositional (3D shapes)	Template	Open QA	✓	✗
CRIC (Gao et al., 2022)	96K	494K	Commonsense (Daily life)	Template	Open QA	✓	✗
AI2D (Hiippala et al., 2021)	5K	15K	Scientific (Science diagram)	Manual	MCQA	✗	✗
ScienceQA (Saikh et al., 2022)	10K	21K	Scientific (Science problems)	Manual	MCQA	✗	✗
MMMU (Yue et al., 2024)	11.5K	11.5K	Scientific (Exam questions)	Manual	MCQA	✗	✗
SEED-Bench (Li et al., 2024)	19K	19K	Commonsense (Spatial, temporal)	Neural	MCQA	✗	✗
MARVEL (Jiang et al., 2024b)	0.8K	3K	Abstract shapes	Template	Open QA	✗	✗
MultiStAR (Direct Answer)	8.1K	21.7K	Abstract shapes	Template & Neural	MCQA	✓	✗
MultiStAR (Logical Chain)	0.56K	3.92K	Abstract shapes	Template & Neural	MCQA	✓	✓

Table 6: Comparison of various VQA datasets. *Template*: generated using predefined rules, *Manual*: written by humans, *Neural*: generated using large language models, *Template & Neural*: generated using predefined rules and rewritten by large language models. *Open QA*: free-text answers, *MCQA*: Multiple-Choice Question Answering. *Functional Program*: Indicates whether the dataset is automatically created by functional programs. *Multi-Step Structure*: Highlights whether the dataset includes a hierarchical structure with interdependent reasoning steps.



(a) Question Length Compare (b) Function Distribution (c) Num. Choice Distribution (d) Prompt Length

Figure 8: The left three panels (a), (b), and (c) present analyses of the Direct Answer task, while the right panel (d) focuses on the Logical Chain task.

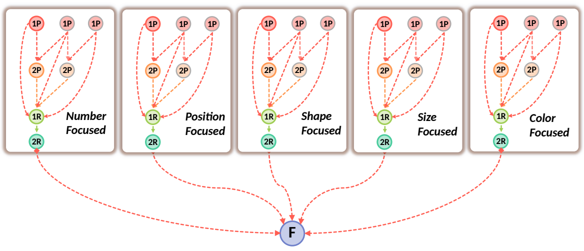


Figure 9: The full logical chain: To arrive at the final answer, we incorporate rules from all five attributes.

To address these settings, when we decompose the problem into subproblems, we treat each part independently. For instance, there are separate question sets for the left and right sections of the panels in the Left-Right setting, with the question explicitly stating which part is being addressed. To clarify the panel structure, an additional sentence is appended to the question:

- **Left-Right:** The panel is divided into two sections by a vertical line, separating the left side from the right side, with objects possibly present in both sections.
- **Up-Down:** The input panel is split by a horizontal line, separating the top side from the bottom side, with objects possibly present in both sections.
- **In-Out:** The panel is divided into two regions: an outer structure and an inner structure, with objects possibly present in both regions.

This extra information is unnecessary for other settings. The complete prompt format is: [Extra Setting Info] **Question:** [question] **Please select one of the following:** [choices]. **The answer should be one of A, B, C, D.**

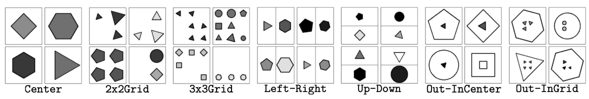


Figure 10: The original RAVEN puzzle, includes seven puzzle settings.

A.4.2 Logical Chain

Model Settings: All MLLMs are tested under their default settings under the environment of Huggingface [2], the *transformer* package version in python is 4.39.2 for NVLM-D-72B model and 4.46.0 for all others. In addition, to handle the

[2] https://huggingface.co/

length of our prompts, we increase the maximum token length to 2048.

When prior information is injected, a rule-based program is used to convert the information into text and integrate it into the prompt. This transformation is necessary because the images referenced in prior questions are not the same as those in the current question, making it impossible to directly reuse them.

For example, if the prior question is "How many objects in this panel?" and the current question is "Comparing the number of objects in the left and right panel," the phrase "this panel" cannot directly correspond to "left panel" or "right panel." To address this, we transform "this panel" into a more specific term, such as "left panel" or "right panel." Table 10 outlines the transfer rules for Number and Position. Similar patterns are applied for other attributes, which are not listed here for brevity.

After the prior information is transformed, the prompt is structured as follows: **[**Extra Setting Info**] Below is the information generated from the previous steps, please be aware that it may or may not contain errors:** [[Prior Info 1], [Prior Info 2], ...] **Question:** [question] **Please select one of the following:** [choices]**. The answer should be one of A, B, C, D.**

A.5 Human Sudies and and Inter-Participant Agreement

To evaluate the subjective quality of human performance in our study, we conducted two separate parts: Part A and Part B. Part A focuses on evaluation of the quality of our automatically generated dataset, while Part B focuses on testing human reasoning abilities across different stages of complexity. For both Part A and Part B, a Consent Form and a Plain Language Statement are provided to the annotators prior to the annotation process. These documents must be read and agreed upon before they can proceed with the annotations.

A.5.1 Part A

Part A involved five research students who participated in answering a series of abstract reasoning questions. This section aimed to evaluate the quality of the dataset generated by our template-based methods. Since the Direct Answer and Logical Chain share the same pool of templates, and Direct Answer covers all templates, we chose to focus on assessing the quality of the Direct Answer component. A random sample of 620 questions was selected for this evaluation. To ensure participants clearly understood the tasks and evaluation criteria, a detailed guide was provided at the beginning of the questionnaire (see Figure 11). An example question from the questionnaire is shown in Figure 12.

Background Information:

You are tasked with reviewing a series of question-answer (QA) pairs related to an abstract visual reasoning dataset. This dataset has been designed to test reasoning abilities across different stages of complexity. The stages range from simple, single-panel reasoning to more complex comparisons between panels.

Dataset Stages for current Task_A:

Basic Perception (Single Panel):

- Focus: Understanding basic properties such as number, position, and shape of objects in a single panel.
- No comparisons are involved in this stage.

Each stage involves different types of reasoning, such as identifying patterns within a single panel or comparing objects across multiple panels. Your feedback will help us evaluate the quality of these questions based on three important criteria: Accuracy, Clarity, and Content Validity.

Explanation of Evaluation Criteria:

1. **Correctness**:
   - This aspect assesses whether the answer provided for the question is correct.
   - **Question**: *Is the provided answer correct for this question?*
   - You will need to evaluate whether the answer corresponds accurately to the visual information presented in the panels.
2. **Clarity**:
   - This aspect evaluates how clear and understandable the question is.
   - **Question**: *Is the question phrased clearly and easy to understand?*
   - Consider if the wording of the question makes sense, whether any terms are ambiguous, and whether someone without prior knowledge could easily comprehend the question.
3. **Content Validity**:
   - This aspect checks if the question is suitable for the task stage in which it is presented.
   - **Question**: *Does the content of the question align with the current task stage?*
   - The dataset is divided into five stages, such as single-panel reasoning or multi-panel comparisons. Evaluate if the question is appropriate for the specific reasoning stage it represents

Instructions for Annotators:

Please carefully review each QA pair and assess it using the three questions provided. Your input will be critical in refining the dataset to ensure it accurately reflects the intended reasoning tasks and provides clear, valid questions and answers.

Figure 11: A detailed guide provided to participants at the beginning of the questionnaire for Part A.

To thoroughly assess the human performance in Part A, we used three key indicators: Correctness, Clarity, and Content Validity.

Correctness assesses whether the answer provided for the question is correct. Evaluators were asked to determine if the provided answer accurately corresponded to the visual information presented in the panels. This involved a careful comparison between the answer and the visual data to ensure accuracy.

Clarity evaluates how clear and understandable the question is. Evaluators considered whether the question was phrased clearly and was easy to understand. They assessed if the wording made sense, if any terms were ambiguous, and whether someone without prior knowledge could easily comprehend the question. This indicator is crucial for ensuring that the questions are accessible and interpretable by all participants.

Content validity checks if the question is suitable for the task stage in which it is presented. Evaluators examined whether the content of the question

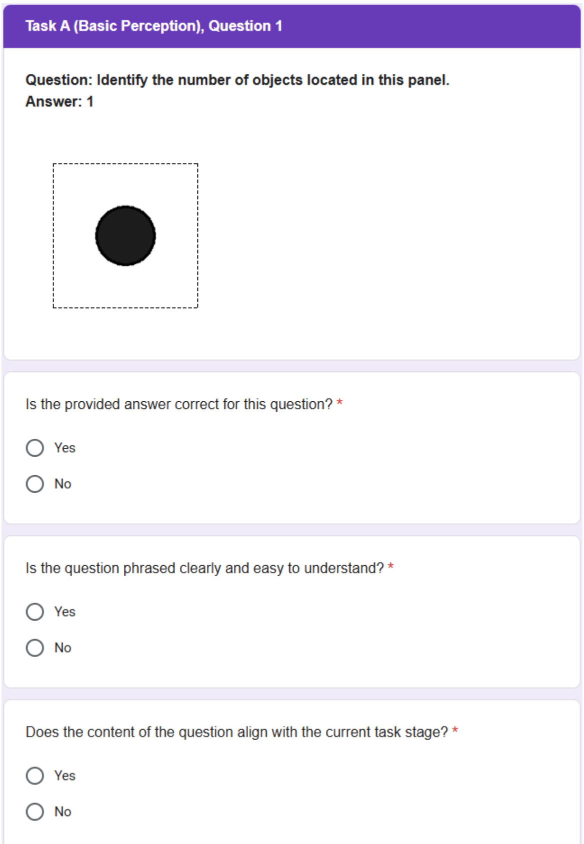


Figure 12: Sample Question from the Questionnaire for Part A.

aligned with the current task stage. The dataset is divided into five types, such as one panel basic perception or two panel comparisons. Participants needed to ensure that the question was appropriate for the specific reasoning type it represented. This indicator ensures that each question is relevant and appropriately challenging for its designated stage.

The metrics used to evaluate performance in Part A included correctness, clarity, and content validity, with positive rates for each metric provided in Table 11. The positive rate is the proportion of questions answered by "Yes". The results indicate that the participants in Part A performed exceptionally well across all metrics, with Correctness, Clarity, and Content Validity scores consistently high. This suggests that the questions were well-designed and comprehensible, and the participants were able to provide accurate answers.

A.5.2 Part B

Part B utilized the Prolific crowdsourcing platform[3] to recruit 162 participants who were subjected to the same set of abstract reasoning questions as those given to the MLLMs. The objective of this part was to evaluate human performance on our dataset, enabling a comparison between human and model capabilities. Participants received a detailed guide at the beginning of the questionnaire, which included task descriptions and several examples, as shown in Figure 13. The guide varied depending on the stage of the Direct Answer task, but for this section, we include only the One-Panel Basic Perception stage. And similar to Part A, due to Direct Answer covering all templates, we chose to focus on assessing the human performance of the Direct Answer component. Each question in the questionnaire for Part B included a image and a multiple-choice question, as illustrated in Figure 14.

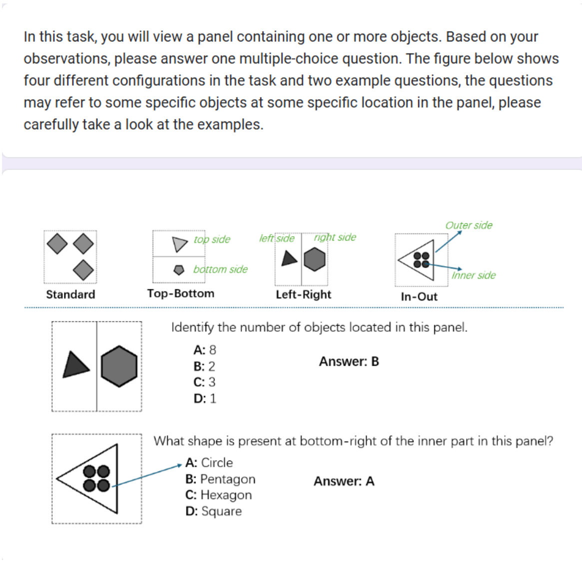


Figure 13: A detailed guide provided to participants at the beginning of the questionnaire for Part B. This guide focuses on One-Panel Basic Perception.

The performance metrics for Part B are summarized in Table 12. The performance for Part B show a noticeable decline in positive rates, particularly for more complex tasks such as Two Panel Compare, One Row, and Two Rows. This decline highlights the increased difficulty of these tasks and suggests that the broader participant pool found these questions more challenging.

Inter-Participant Agreement. To quantify the inter-participant agreement across participants for Part B stuides, we computed Fleiss' kappa scores (Landis and Koch, 1977) across different question types. The Fleiss' Kappa scores for each question types are provided in Table 13.

The high Fleiss' Kappa score for One-Panel Ba-

[3] https://www.prolific.com/

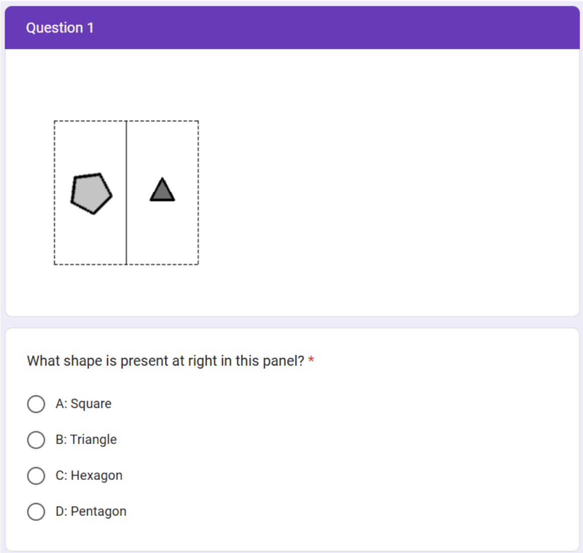


Figure 14: Sample Question from the One-Panel Basic Perception Questionnaire for Part B.

sic (0.9711) indicates strong agreement among the participants, this is mainly due to the simplicity of One-Panel Basic questions. However, the lower scores for Two-Panel and rule deduction phase highlight the increased difficulty and the significant variability in human interpretation for these more complex tasks.

A.6 Performance Increase with Prior Info

Tables 15 and 16 present the percentage increase in Accuracy and MSEval metrics, respectively. It is evident that, except for the Final stage, all other stages show improved performance, with MSEval and Accuracy metrics closely aligned in these cases. However, in the Final stage, while Accuracy does not show a significant increase for the four open-source models, MSEval suggests some improvement due to the incorporation of rule information for solving the final RAVEN puzzle. An exception is observed with Qwen2-VL-72B, which may already perform well on RAVEN. Incorporating information from earlier stages might introduce misleading details, leading to a significant performance drop.

A.7 Additional Details about MSEval

A.7.1 Algorithm Pseudo Code

Algorithm 1 shows the details Pseudo Code for our proposed MSEval metrics.

A.7.2 Computational Cost

The computational complexity of the algorithm can be expressed as:

$$O(N \cdot |\mathcal{E}_t| \cdot |A|)$$

where:

- N: The number of samples.
- $|\mathcal{E}_t|$: The number of edges in the logical chain (dependency relationships between nodes).
- $|A|$: The number of possible choices for each node.

In most cases, $|A|$, the number of possible choices per node, is typically equal to 4. As a result, the computational complexity simplifies to:

$$O(4 \cdot N \cdot |\mathcal{E}_t|) \quad \text{or simply} \quad O(N \cdot |\mathcal{E}_t|),$$

which is effectively linear with respect to both the number of instances (N) and the number of edges ($|\mathcal{E}_t|$) in the logical chain. By reducing the number of edges or employing a smaller logical chain, the computational cost can be significantly minimized, ensuring better scalability and efficiency, especially for large datasets or complex logical dependencies. This simplification highlights the importance of optimizing the chain structure to maintain computational feasibility. The actual time cost for each open-source model we tested is shown in Table 14.

A.8 Discussion Additional Materials

A.8.1 Model Parameters Trend

Figures 17, 18, 19, 20, 21, 22, and 23 demonstrate that performance generally improves with larger model parameter sizes across stages, except for the 2R and Final stages. For these two stages, most models perform below the random baseline. The differences in model performance are primarily attributed to the varying sizes of their language encoders, highlighting the significant role of a robust language encoder in overall performance. However, despite the observed improvements, a noticeable gap persists between model and human performance. This discrepancy may arise from the complexity of the visual input, which poses challenges for models in fully understanding and integrating multimodal information.

A.8.2 Attribute Break-Down Analysis

Figure 24 illustrates the attribute-level performance breakdown of two open-source models and four closed-source models evaluated on the logical chain task. Gemini, GPT-4o, and the two larger

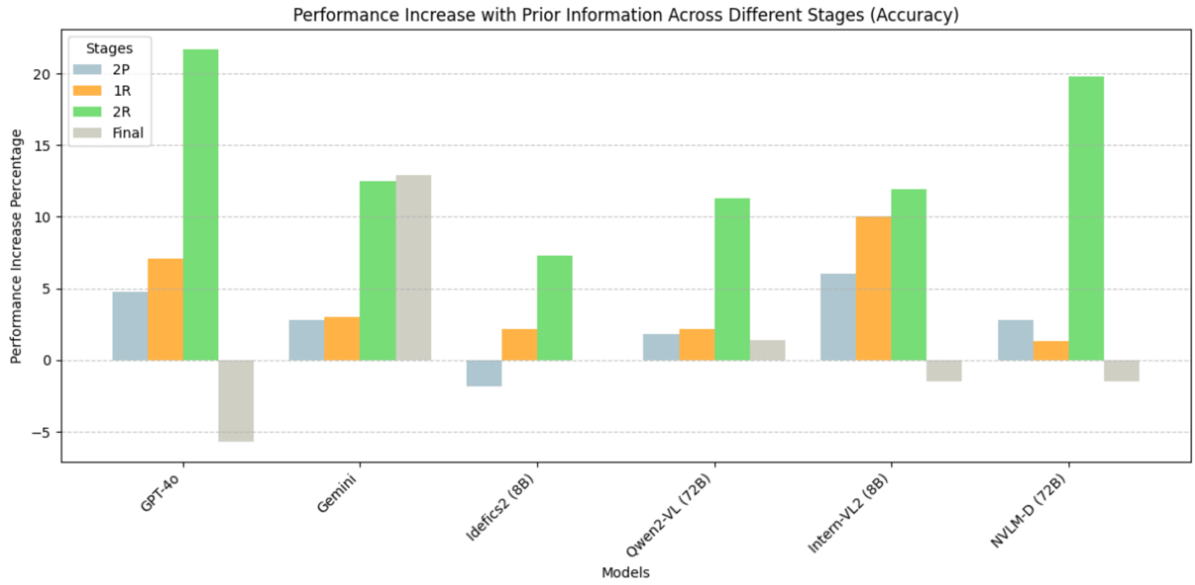


Figure 15: Accuracy percentage increase after incorporating prior info.

models, Qwen2-VL and NVLM-D, exhibit similar trends: the Number attribute achieves the highest performance in more complex stages (2P, 1R, and 2R), while Position dominates in lower-level stages (1P-C and 1P-B). In contrast, smaller models like Idefics2 and Intern2-VL struggle with the Number attribute but perform relatively better on Position, indicating that these models are less sensitive to counting tasks but demonstrate better spatial reasoning.

A.8.3 Handling Long Prompts

Table 15 presents the accuracy and MSEval scores for three prompting techniques incorporating prior information in the Logical Chain task. For GPT-4o, Qwen2-VL, and Gemini, the use of HTML tags yields significant performance improvements. Additionally, for GPT-4o and Qwen2-VL, the Document-based prompting also demonstrates notable benefits. However, for other models, these two techniques show a negative impact. In this case, the MSEval results are consistent with the accuracy outcomes. List 1 shows an example of HTML structured prompts, while Figure 25 shows an example of Document structured prompts.

```
<!DOCTYPE html>
<html>
  <body>
    <h1>
      In this visual puzzle, you are
          given two panels. Each panel
          divided into
      two sections by a vertical line,
          separating the <strong>left</
          strong>
      side from the <strong>right</
          strong> side, with objects
          might present
      in both sections. Below is the
          information generated from the
           previous steps,
      please be aware that it may or may
           not contain errors:
    </h1>
    <div>
      <h2>Panel Information</h2>
      <ul>
        <li>There are 1 objects in the <
            strong>left</strong> part of
             the
        <strong>left</strong> panel.</li
            >
        <li>There are 2 objects in the <
            strong>left</strong> part of
             the
        <strong>right</strong> panel.</
            li>
      </ul>
    </div>
    <div>
      <h2>Question</h2>
      <p>
        Consider only the <strong>left</
            strong> part of the two
            panels in the image.
        Does the <strong>left</strong>
            panel contain the same
            number of objects,
        more objects, or fewer objects
            than the <strong>right</
            strong> panel?
        Please select one of the
            following:
      </p>
      <ul>
        <li>A: More</li>
        <li>B: The same</li>
        <li>C: Fewer</li>
```

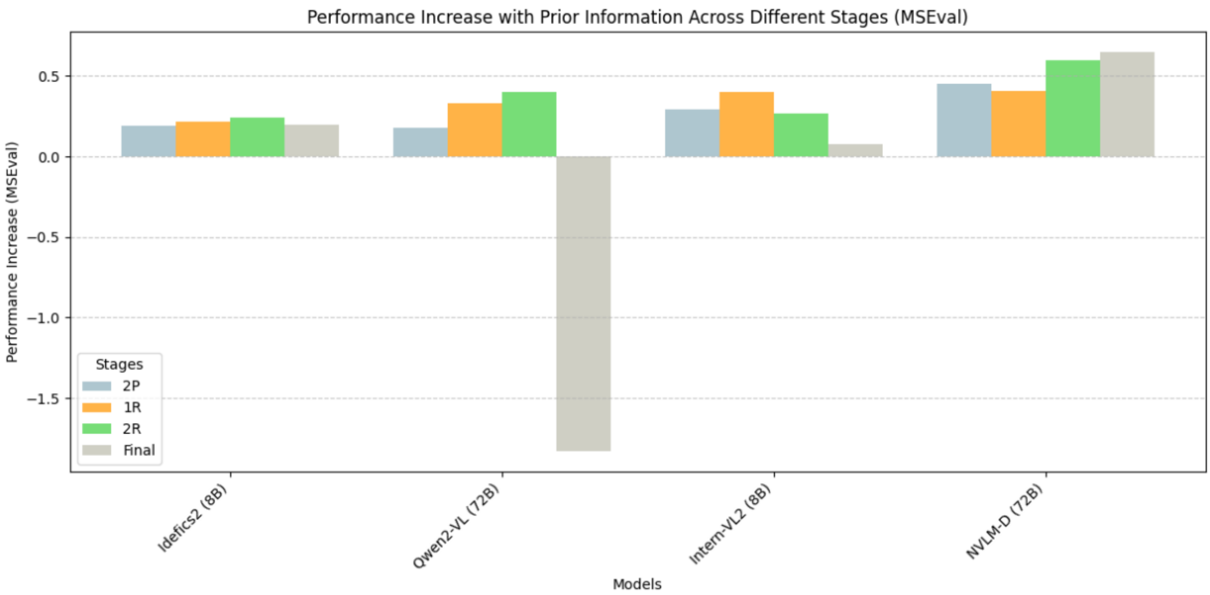


Figure 16: MSEval percentage increase after incorporating prior info.

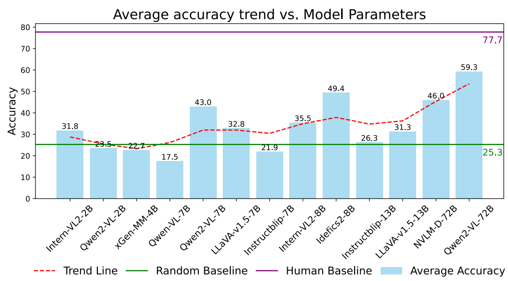


Figure 17: The average accuracy trend for the Direct Answer task as model sizes gradually increase. The trend line is derived using Gaussian smoothing, and the average accuracy is calculated by averaging the results across all five stages.

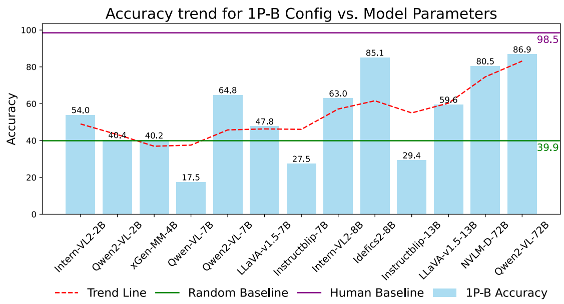


Figure 18: The One-Panel Basic Perception accuracy trend for the Direct Answer task as model sizes gradually increase. The trend line is derived using Gaussian smoothing.

```
            </ul>
            <p>The answer should be one of A,
                B, C.</p>
        </div>
    </body>
</html>
```

Listing 1: HTML Structure for Handling Long Prompt

A.9 Error Analysis

As shown in Figure 26, we provide an example for each type of error:

- **Perception Error:** This occurs when the model misinterprets visual inputs, such as object numbers or shapes. In the provided example, there should be four objects in the left panel and three in the right panel, but the model fails to recognize this correctly.
- **Reasoning Error:** This involves incorrect logic applied to correctly perceived inputs. In this example, the model accurately identifies the objects and their edge numbers in each panel. However, due to flawed reasoning, it incorrectly concludes that the number of edges is decreasing, which is not the case.
- **Hallucination Error:** This error type refers to the generation of incomplete sentences or unrelated information. Here, the question asks about the position of objects, but the explanation provided by the model focuses on shapes and sizes, which are irrelevant.
- **Propagation Error:** This occurs when the model fails to detect or correct inaccuracies in prior information. In this example, the prior information is already incorrect, but the model

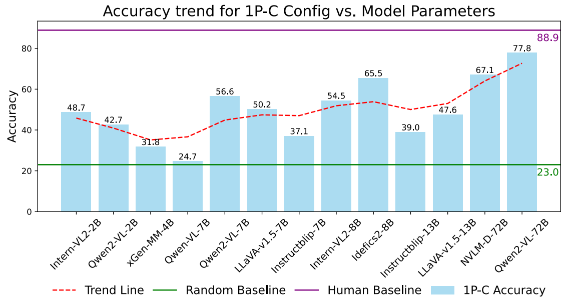


Figure 19: The One-Panel Comparison accuracy trend for the Direct Answer task as model sizes gradually increase. The trend line is derived using Gaussian smoothing.

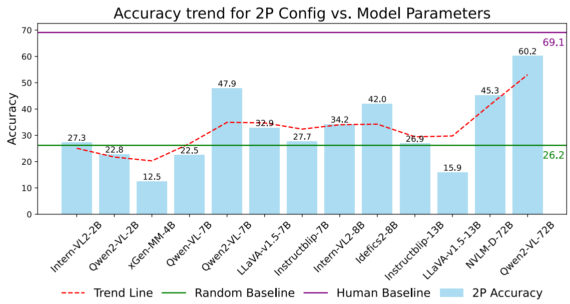


Figure 20: The Two-Panels Comparison accuracy trend for the Direct Answer task as model sizes gradually increase. The trend line is derived using Gaussian smoothing.

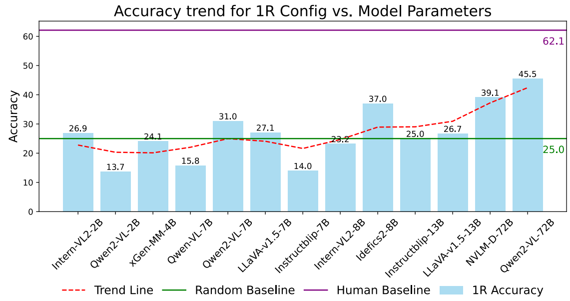


Figure 21: The One-Row Deduction accuracy trend for the Direct Answer task as model sizes gradually increase. The trend line is derived using Gaussian smoothing.

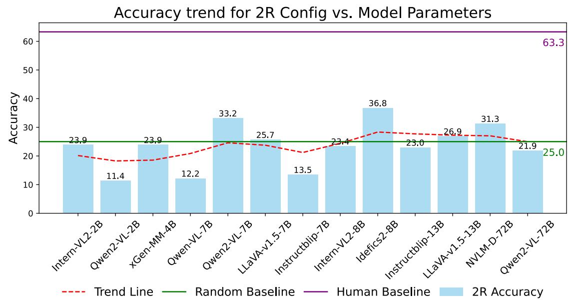


Figure 22: The Two-Rows Deduction accuracy trend for the Direct Answer task as model sizes gradually increase. The trend line is derived using Gaussian smoothing.

does not identify or address these inaccuracies, leading to an incorrect answer.

A.10 Qualitative Analysis

Lists 2 to 6 present five examples that illustrate the advantages of our MSEval metric over traditional accuracy. For instance, in List 5, all the intermediate steps are incorrect, yet the model arrives at the correct answer with only a small confidence margin (31.375 compared to the second-highest confidence of 31.125). In this case, the model's performance is not truly effective, as it is unclear how it managed to produce the correct answer despite incorrect intermediate steps. Unlike traditional accuracy, which would mark this as fully correct, MSEval appropriately penalizes such cases by assigning a low score.

Conversely, a reverse scenario is shown in List 4, where traditional accuracy marks the result as entirely incorrect. However, the model correctly answers all intermediate steps, and the probability of the correct answer is very close to the highest confidence value. In this situation, MSEval assigns a relatively high score, reflecting the model's partial success and rewarding its correct reasoning process.

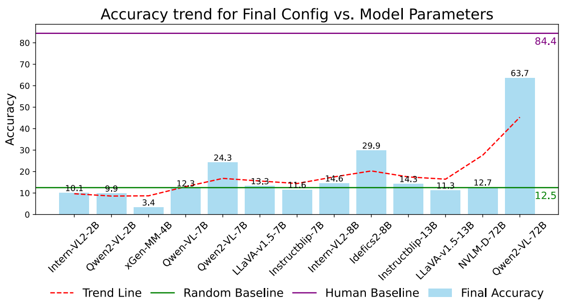


Figure 23: The Final accuracy trend for the Direct Answer task as model sizes gradually increase. The trend line is derived using Gaussian smoothing.

Algorithm 1 Overall Workflow

Input:

Logical Chain $\mathcal{D}_t$

Define: $\mathcal{S}_t = \{t\} \cup \mathcal{D}_t$

Model logits $\mathcal{Z} = \{z_j^{(i)} \mid j \in \mathcal{S}_t, i = 1, \dots, N\}$

All possible choices for each node $\{\mathcal{A}_j^{(i)} \mid j \in \mathcal{S}_t\}$

Output: MSEval score for stage t: MSEval_t

Step 1: Compute Conditional Probabilities

for each sample $i = 1$ to N **do**

  for each node $j \in \mathcal{S}_t$ **do**

    Compute probability $p_j^{(i)} \leftarrow \frac{\exp(z_j^{(i)})}{\sum_{k \in \mathcal{A}_j^{(i)}} \exp(z_k^{(i)})}$

  end for

end for

Step 2: Compute Conditional Mutual Information

for each sample $i = 1$ to N **do**

  for each node $j \in \mathcal{S}_t$ **do**

    Alter $\mathcal{A}_j^{(i)}$ to generate perturbed outputs $\mathcal{A}_{j \to t}^{(i)}$

    Compute:

$$\begin{aligned}\text{CMI}(i, j, t) \leftarrow\; & H(\mathcal{A}_{j \to t}^{(i)} \mid \mathcal{D}_t^{(i),-j}) \\ & + H(\mathcal{A}_j^{(i)} \mid \mathcal{D}_t^{(i),-j}) \\ & - H(\mathcal{A}_{j \to t}^{(i)}, \mathcal{A}_j^{(i)} \mid \mathcal{D}_t^{(i),-j})\end{aligned}$$

  end for

end for

Step 3: Normalize Conditional Mutual Information

for each sample $i = 1$ to N **do**

  for each node $j \in \mathcal{S}_t$ **do**

    Compute:

$$\text{NCMI}(i, j, t) \leftarrow \frac{\exp(\text{CMI}(i, j, t))}{\sum_{k \in \mathcal{S}_t} \exp(\text{CMI}(i, k, t))}$$

  end for

end for

Step 4: Compute MSEval Score for Each Sample

for each sample $i = 1$ to N **do**

  Initialize $\text{MSEval}_t^{(i)} \leftarrow 0$

  for each node $j \in \mathcal{S}_t$ **do**

    Compute

$$\epsilon_j^{(i)} = \frac{1}{|\mathcal{A}_j^{(i)}|}$$

    Update:

$$\text{MSEval}_t^{(i)} \leftarrow \text{MSEval}_t^{(i)} + \text{NCMI}(i, j, t) \cdot \frac{p_j^{(i)}}{\epsilon_j^{(i)}}$$

  end for

end for

Step 5: Average MSEval Across All Samples

Compute:

$$\text{MSEval}_t \leftarrow \frac{1}{N} \sum_{i=1}^{N} \text{MSEval}_t^{(i)}$$

Return: MSEval_t

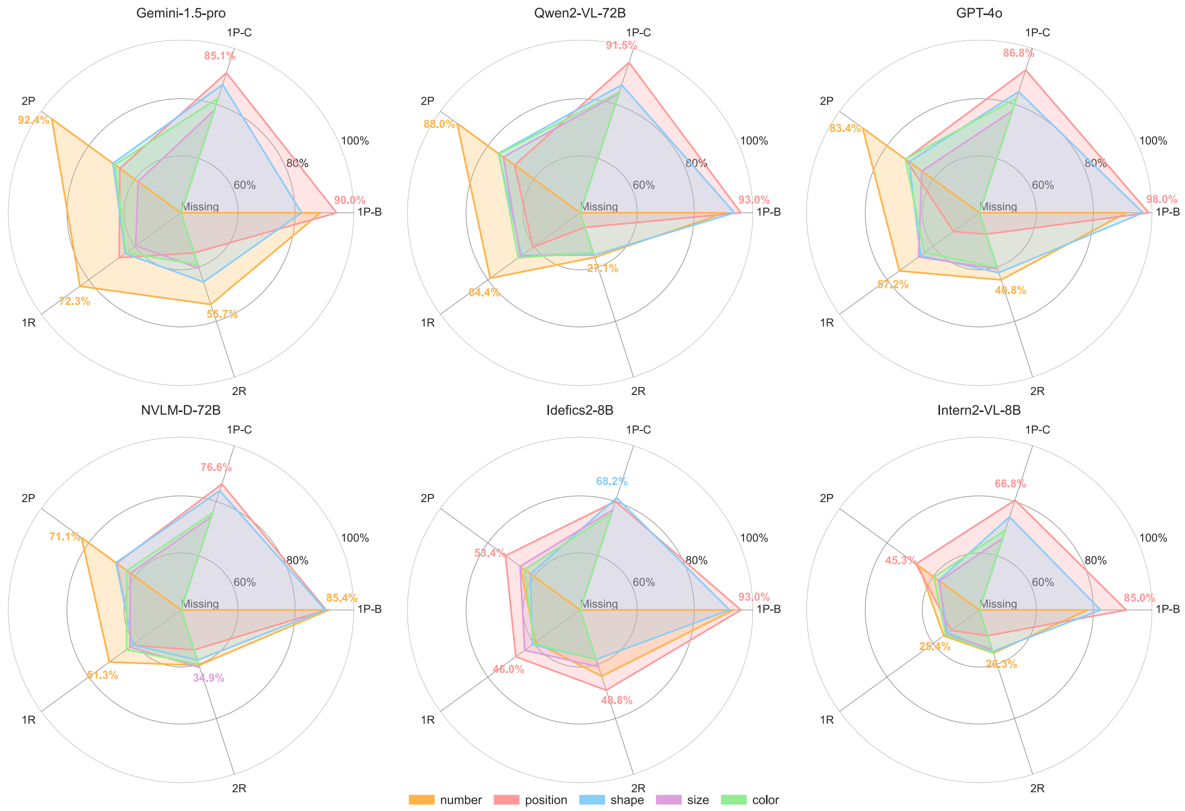
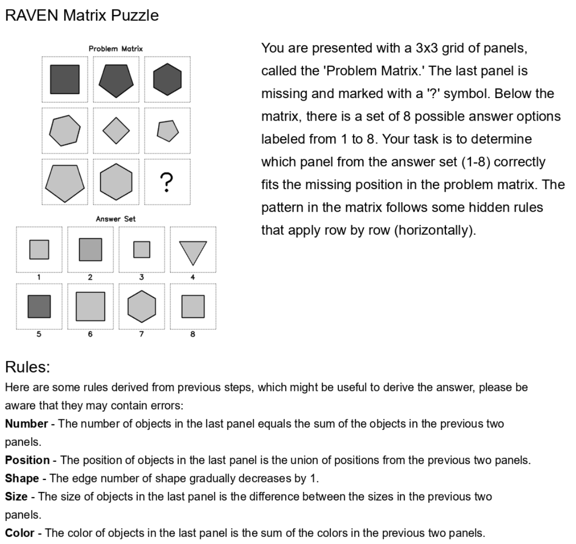


Figure 24: The Radar Chart for all six models in Logical Chain subtask.


RAVEN Matrix Puzzle

Problem Matrix

Answer Set

You are presented with a 3x3 grid of panels, called the 'Problem Matrix.' The last panel is missing and marked with a '?' symbol. Below the matrix, there is a set of 8 possible answer options labeled from 1 to 8. Your task is to determine which panel from the answer set (1-8) correctly fits the missing position in the problem matrix. The pattern in the matrix follows some hidden rules that apply row by row (horizontally).

Rules:

Here are some rules derived from previous steps, which might be useful to derive the answer, please be aware that they may contain errors:

Number - The number of objects in the last panel equals the sum of the objects in the previous two panels.

Position - The position of objects in the last panel is the union of positions from the previous two panels.

Shape - The edge number of shape gradually decreases by 1.

Size - The size of objects in the last panel is the difference between the sizes in the previous two panels.

Color - The color of objects in the last panel is the sum of the colors in the previous two panels.

Figure 25: Document Structure for Handling Long Prompt.

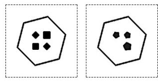
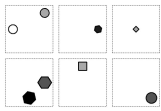
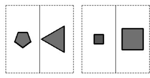
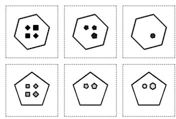
Error Type	Question	GT Answer	Generated Answer + Explanation	Image
Perception Error	Consider only the inner part of the two panels in the image. Does the left panel contain the same number of objects, more objects, or fewer objects than the right panel?	More Objects	Fewer Objects. Explanation: In the inner part of the left panel, there are three objects. In the inner part of the right panel, there are four objects. Therefore, the left panel contains fewer objects than the right panel.	
Reasoning Error	Examine the three panels in the first row, then the three panels in the second row, both from left to right, and derive a rule that applies to both rows in relation to the shape of objects.	Three distinct shapes across panels, rotating through each possible permutation of these shapes.	The edge number of shape gradually decreases by 1. Explanation: Panel 1: Circle (infinite edges) Panel 2: Hexagon (6 edges) Panel 3: Quadrilateral (4 edges) Panel 1: Hexagon (6 edges) Panel 2: Quadrilateral (4 edges) Panel 3: Circle (infinite) In both rows, the number of edges of the shape decreases as we move from left to right.	
Hallucination Error	Consider only the left part of the two panels in the image. Is the position of all the objects in the left panel the same as the objects in the right panel?	Yes	No The left part of the left panel contains a pentagon. The left part of the right panel contains a small square. The shapes and their sizes are different.	
Propagation Error	Analyze the first row of three panels from left to right, followed by the second row of three panels, concentrating only on the **outer** sections of each panel, and determine a rule applicable to both rows that dictates the position of objects. Prior Info: If an object is in the first panel but not in the second at corresponding position, it appears in the third panel.	The position of objects does not change across panels.	If an object is in the first panel but not in the second at corresponding position, it appears in the third panel. Explanation: The pattern observed is that if an object (in this case, a black square) is present in the first panel but not in the second panel at the corresponding position, it appears in the third panel. This rule is consistent across both rows.	

Figure 26: Examples for four types of error.

Question Pattern	Question Example	Attribute	Constraints	Answer Space
One-Panel Basic				
How many objects are in panel <P>?	How many objects are in panel 1?	Number	NA	[1,2,3,4,5,6,7,8,9]
What is the shape of the object at <X> in panel <P>?	What is the shape of the object at top-left in panel 1?	Shape	NA	["triangle", "square", "pentagon", "hexagon", "circle"]
Where is the <S> positioned in panel <P>?	Where is the triangle positioned in panel 1?	Position	NA	["Left", "Right", "Top", "Down", "Bottom-Left", ...]
One-Panel Comparison				
In panel <P>, is the shape of the object on the <X> have the same, more, or fewer edges compared to the object on the <X2>? (Note: The edge number increases in the following order: triangle, square, pentagon, hexagon, circle)	In this panel, is the shape of the object on the left have the same, more, or fewer edges compared to the object on the right? (Note: The edge number increases in the following order: triangle, square, pentagon, hexagon, circle)"	Shape	Not_Equal(X, X2)	["The same", "Fewer", "More"]
In panel <P>, does the object on the <X> the same, smaller or larger in size compared to the object on the <X2>?	In panel 1, does the object on the top-left the same, smaller or larger in size compared to the object on the bottom-right?	Size	Not_Equal(X, X2)	["The same", "Smaller", "Larger"]
In panel <P>, does the object on the <X> the same, darker or brighter in color compared to the object on the <X2>?	In panel 1, does the object on the top-left the same, darker or brighter in color compared to the object on the bottom-right?	color	Not_Equal(X, X2)	["The same", "Darker", "Brighter"]
In panel <P>, where is the <S> relative to the <S2>?	In panel 1, where is the triangle relative to the square?	Position	Not_Equal(S, S2)	["Left", "Right", "Above", "Below", ...]
Are all objects in panel <P> of the same shape?	Are all objects in panel 1 of the same shape?	Shape	NA	["Yes", "No"]
Are all objects in panel <P> of the same size?	Are all objects in panel 1 of the same size?	Size	NA	["Yes", "No"]
Are all objects in panel <P> of the same color?	Are all objects in panel 1 of the same color?	Color	NA	["Yes", "No"]
Two-Panels Comparison				
Does panel <P> contain the same number of objects, more objects, or fewer objects than panel <P2>?	Does panel 1 contain the same number of objects, more objects, or fewer objects than panel 2?	Number	Not_Equal(P, P2), Same_Row(P, P2)	["The same", "More", "Fewer"]
Is the shape of all the objects in panel <P> have the same, more, or fewer edges compared to the objects in panel <P2>? If the shapes within either panel are already different from each other, select 'Not Comparable.' (Note: The edge number increases in the following order: triangle, square, pentagon, hexagon, circle)	Is the shape of all the objects in panel 1 have the same, more, or fewer edges compared to the objects in panel 2? If the shapes within either panel are already different from each other, select 'Not Comparable.' (Note: The edge number increases in the following order: triangle, square, pentagon, hexagon, circle)	Shape	Not_Equal(P, P2), Same_Row(P, P2)	["The same", "Fewer", "More", "Not comparable"]
Is the size of all the objects in panel <P> the same as, smaller or larger than the objects in panel <P2>? If the sizes within either panel are already different from each other, select 'Not Comparable.'	Is the size of all the objects in panel 1 the same as, smaller or larger than the objects in panel 2? If the sizes within either panel are already different from each other, select 'Not Comparable.'	Size	Not_Equal(P, P2), Same_Row(P, P2)	["The same", "Smaller", "Larger", "Not comparable"]

Question Pattern	Question Example	Attribute	Constraints	Answer Space
Two-Panels Comparison				
Is the color of all the objects in panel <P> the same as, darker or brighter than the objects in panel <P2>? If the colors within either panel are already different from each other, select 'Not Comparable.'	Is the color of all the objects in panel 1 the same as, darker or brighter than the objects in panel 2? If the colors within either panel are already different from each other, select 'Not Comparable.'	Color	Not_Equal(P, P2), Same_Row(P, P2)	["The same", "Darker", "Brighter", "Not comparable"]
Is the position of all the objects in panel <P> the same as the objects in panel <P2>?	Is the position of all the objects in panel 1 the same as the objects in panel 2?	Position	Not_Equal(P, P2), Same_Row(P, P2)	["Yes", "No"]
One-Row Rule Deduction				
Examine the three panels in the image from left to right and identify the rule that governs the number of the objects.	-	Number	NA	See Number Rule in Table 8
Examine the three panels in the image from left to right and identify the rule that governs the position of the objects.	-	Position	NA	See Position Rule in Table 8
Examine the three panels in the image from left to right and identify the rule that governs the shape of the objects.	-	Shape	NA	See Shape Rule in Table 8
Examine the three panels in the image from left to right and identify the rule that governs the size of the objects.	-	Size	NA	See Size Rule in Table 8
Examine the three panels in the image from left to right and identify the rule that governs the color of the objects.	-	Color	NA	See Color Rule in Table 8
Two-Rows Rule Deduction				
Inspect the first row of three panels from left to right and inspect the second row of three panels from left to right and determine a rule applicable to both rows that governs the number of objects.	-	Number	NA	Same as One-Row
Inspect the first row of three panels from left to right and inspect the second row of three panels from left to right and determine a rule applicable to both rows that governs the position of objects.	-	Position	NA	Same as One-Row
Inspect the first row of three panels from left to right and inspect the second row of three panels from left to right and determine a rule applicable to both rows that governs the shape of objects.	-	Shape	NA	Same as One-Row
Inspect the first row of three panels from left to right and inspect the second row of three panels from left to right and determine a rule applicable to both rows that governs the size of objects.	-	Size	NA	Same as One-Row
Inspect the first row of three panels from left to right and inspect the second row of three panels from left to right and determine a rule applicable to both rows that governs the color of objects.	-	Color	NA	Same as One-Row

Table 7: Question pattern templates with corresponding example questions. There are 25 templates in total. 3 templates for One-Panel Basic. 7 templates for One-Panel Comparison. 5 templates for Two-Panels Comparison. 5 Templates for One-Row Rule Deduction. 5 Templates for Two-Rows Rule Deduction.

Category	Value Ranges
	Placeholders
Position (<X>)	center, left, right, top, bottom, top-left, top-right, bottom-left, bottom-right, top-left, top-center, top-right, middle-left, middle-center, middle-right, bottom-left, bottom-center, bottom-right, outer-part, inner-part, top-left of the inner part, top-right of the inner part, bottom-left of the inner part, bottom-right of the inner part
Panel (<P>)	0, 1, 2, 3, 4, 5, 6, 7
Shape (<S>)	triangle, square, pentagon, hexagon, circle
	Rules
Number Rule	The number of objects gradually decreases by 1; The number of objects remains constant; The number of objects gradually increases by 1; The number of objects distributes three distinct values across panels, rotating through each possible permutation of these values; The number of objects in the last panel equals the sum of the objects in the previous two panels; The number of objects in the last panel equals the difference between the objects in the previous two panels; No clear rule is present
Position Rule	If an object is in the first panel but not in the second at corresponding position, it appears in the third panel; The position of objects in the last panel is the union of positions from the previous two panels; Three distinct position settings across panels, rotating through each possible permutation of these settings; The position of objects does not change across panels; No clear rule is present
Color Rule	The color of objects gradually darkens by a constant amount each time; The color of objects gradually brightens by a constant amount each time; The color of objects in the last panel is the sum of the colors in the previous two panels; The color of objects in the last panel is the difference between the colors in the previous two panels; Three distinct colors across panels, rotating through each possible permutation of these colors; The color remains constant; No clear rule is present
Size Rule	The size of objects gradually increases by a constant amount each time; The size of objects gradually decreases by a constant amount each time; The size of objects in the last panel is the sum of the sizes in the previous two panels; The size of objects in the last panel is the difference between the sizes in the previous two panels; Three distinct sizes across panels, rotating through each possible permutation of these sizes; The size remains constant; No clear rule is present
Shape Rule	The edge number of shape gradually decreases by 1; The edge number of shape gradually increases by 1; Three distinct shapes across panels, rotating through each possible permutation of these shapes; The shape remains constant; No clear rule is present

Table 8: Pre-defined placeholder value ranges and rules for five attributes

Attributes	Logical Chain Stage Example
Number	1P: How many objects are in the panel? 2P: Does the left panel contain the same number of objects, more objects, or fewer objects than the right panel? 1R: Inspect the three panels in the image from left to right and identify the rule that dictates the number of objects. 2R: Inspect the first row of three panels from left to right and inspect the second row of three panels from left to right and determine a rule applicable to both rows that governs the number of objects.
Position	1P: Where is the circle positioned in the panel? 2P: Is the position of all the objects in the left panel the same as the objects in the right panel? 1R: Examine the three panels in the image from left to right and identify the rule that governs the position of the objects. 2R: Examine the three panels in the first row, then the three panels in the second row, both from left to right, and derive a rule that applies to both rows in relation to the position of objects.
Shape	1P: What is the shape of the object at center in the panel? 2P: Is the shape of all the objects in the left panel have the same, more, or fewer edges compared to the objects in the right panel? 1R: Inspect the three panels in the image from left to right and identify the rule that dictates the shape of objects. 2R: Analyze the first row of three panels from left to right, followed by the second row of three panels, and identify a common rule that dictates the shape of objects in both rows.
Size	1P: Are all objects in the panel of the same size? 2P: Is the size of all the objects in the left panel the same as, smaller or larger than the objects in the right panel? 1R: Analyze the three panels in the image from left to right and uncover the rule that governs the size of objects. 2R: Review the first row of three panels in sequence from left to right, then do the same for the second row, and determine a shared rule that governs the size of objects in both rows.
Color	1P: Are all objects in the panel of the same color? 2P: Is the color of all the objects in panel <P> the same as, darker or brighter than the objects in panel <P2>? 1R: Inspect the three panels in the image from left to right and identify the rule that dictates the color of objects. 2R: Examine the three panels in the first row, then the three panels in the second row, both from left to right, and derive a rule that applies to both rows in relation to the color of objects.
Final	You are presented with a 3x3 grid of panels, called the *Problem Matrix*. The last panel is missing and marked with a '?' symbol. Below the matrix, there is a set of 8 possible answer options labeled from 1 to 8. Your task is to determine which panel from the answer set (1-8) correctly fits the missing position in the problem matrix. The pattern in the matrix follows some hidden rules that apply row by row (horizontally). Please select the number (from 1 to 8) of the panel that completes the pattern.

Table 9: A full logical chain with the examples for five stages.

Attribute	Stage	Numbers	Output
Number	`single_panel`	`["1"]`	There are `{answer_str}` objects in the left panel.
		`["2"]`	There are `{answer_str}` objects in the right panel.
		`["3"]`	There are `{answer_str}` objects in the right panel.
	`two_panels`	`["1", "2"]`	The left panel has `{answer_str}` objects compared to the middle panel.
		`["2", "3"]`	The middle panel has `{answer_str}` objects compared to the right panel.
	`one_row`	Any	The rule for the number of objects in the first row is: `{answer_str}`.
Position	`single_panel`	`["1"]`	`Where is the (\w+) positioned in the panel?` becomes: `There is a \1 positioned in the left panel.`
		`["2"], ["3"]`	`Where is the (\w+) positioned in the panel?` becomes: `There is a \1 positioned in the right panel.`
	`two_panels`	`["1", "2"]`	If `answer_str` is `Yes`, "The position of all the objects in the left panel is the same as the objects in the middle panel." Otherwise, "The position of all the objects in the left panel is not the same as the objects in the middle panel."
		`["2", "3"]`	If `answer_str` is `Yes`, "The position of all the objects in the middle panel is the same as the objects in the right panel." Otherwise, "The position of all the objects in the middle panel is not the same as the objects in the right panel."
	`one_row`	Any	The rule for the position of objects in the first row is: `{answer_str}`.

Table 10: Rule-based program of attribute Number and Position. The Stage represents the prior stage. (\w+) represents the word here will be put in the position of \1.

Metric	One Panel Basic	One Panel Compare	Two Panel Compare	One Row	Two Rows
Correctness	0.98	0.97	0.96	0.93	0.94
Clarity	0.96	0.97	0.94	0.95	0.99
Content Validity	0.99	0.99	0.99	1.00	1.00

Table 11: Human performance (positive rates) for Part A across different question types.

1P-B	1P-C	2P	1R	2R
98.52	88.89	69.08	62.12	63.33

Table 12: Human performance (positive rates) for Part B across different question types.

Task	1P-B	1P-C	2P	1R	2R
Kappa Scores	0.9711	0.7830	0.4988	0.4443	0.4075

Table 13: Fleiss' Kappa Scores for Inter-Participant Agreement across different question types.

Model	Running Time
Idefics2-8B	7H 24M
Intern2-VL-8B	14H 54M
Qwen2-VL-Instruct-72B	5D 3H 50M
NVLM-D-72B	5D 0H 13M

Total Questions: 3.92K

Device: 2 × A100 80G GPU

Table 14: Actual Running Time for Each Model. D: Day, H: Hour, M: Minute

	Metric	Prior	1P	2P	1R	2R	Final
GPT-4o	Acc	Vanilla	73.8	43.9	41.8	50.6	10.0
		Struct.	**82.2**	64.4	47.8	50.9	8.6
		Doc.	80.8	44.8	31.1	24.9	10.0
Gemini	Acc	Vanilla	75.5	64.4	52.6	57.1	18.6
		Struct.	70.6	66.4	52.9	**57.8**	17.1
		Doc.	69.6	51.0	36.7	33.1	14.3
Qwen2-VL (72B)	Acc	Vanilla	74.1	57.8	47.3	54.2	65.7*
		Struct.	77.2	**67.7**	*55.1*	53.6	61.4
		Doc.	76.5	63.1	50.2	46.6	24.3
	MSEval	Vanilla	2.54	1.95	1.79	1.70	5.14*
		Struct.	2.64	2.46	2.37	2.17	3.11
		Doc.	2.62	2.33	2.26	1.90	1.88
NVLM-D (72B)	Acc	Vanilla	66.1	45.2	39.1	43.3	7.1
		Struct.	45.6	25.3	23.1	36.8	10.0
		Doc.	18.7	11.3	17.5	14.2	20.0
	MSEval	Vanilla	2.25	1.20	1.28	1.02	0.76
		Struct.	1.84	0.87	0.99	0.98	0.79
		Doc.	0.93	0.78	0.88	0.94	0.96
Idefics2 (8B)	Acc	Vanilla	57.8	37.8	36.6	42.4	25.7
		Struct.	46.2	30.5	36.9	43.5	18.6
		Doc.	27.2	23.6	9.6	6.9	15.7
	MSEval	Vanilla	2.02	1.48	1.51	1.51	1.44
		Struct.	1.59	1.18	1.25	1.33	1.27
		Doc.	1.04	1.01	0.97	0.98	1.00
Intern2-VL (8B)	Acc	Vanilla	54.4	41.9	31.6	33.5	17.1
		Struct.	48.4	31.0	20.7	27.3	7.1
		Doc.	23.1	30.2	17.1	17.1	8.6
	MSEval	Vanilla	2.02	1.48	1.51	1.51	1.44
		Struct.	1.52	1.18	1.10	1.00	0.97
		Doc.	1.02	1.00	1.00	0.97	0.92

Table 15: The Accuracy (Acc) and MSEval scores of three prompting techniques for the Logical Chain task. Vanilla: Pure Text, Struct.: Structure (HTML), Doc.: Document. The highest accuracy are highlighted in **bold**. The highest MSEval are highlighted in underline.

```
--------------------------------
Dependent Stage Name: single_panel_1_left
Dependent Stage Question: Are all objects in the left part of the panel of the same
color?
Dependent Stage Choice: ['A: Only one object', 'B: No', 'C: Yes']
Dependent Stage Ground Truth: A
Dependent Stage Logits: {'A': 11.125, 'B': 11.125, 'C': 20.75}
Dependent Stage Generated Answer: C
--------------------------------
Dependent Stage Name: single_panel_2_left
Dependent Stage Question: Are all objects in the left part of the panel of the same
color?
Dependent Stage Choice: ['A: No', 'B: Only one object', 'C: Yes']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 11.125, 'B': 10.5625, 'C': 19.875}
Dependent Stage Generated Answer: C
--------------------------------
Dependent Stage Name: single_panel_3_left
Dependent Stage Question: Are all objects in the left part of the panel of the same
color?
Dependent Stage Choice: ['A: Only one object', 'B: No', 'C: Yes']
Dependent Stage Ground Truth: A
Dependent Stage Logits: {'A': 10.375, 'B': 10.375, 'C': 19.75}
Dependent Stage Generated Answer: C
--------------------------------
Dependent Stage Name: two_panels_1_2_left
Dependent Stage Question: Consider only the left part of the two panels in the image.
Is the color of all the objects in the left panel the same as, darker or brighter than
the objects in the right panel? If the colors within either panel are already different
 from each other, select 'Not Comparable.'
Dependent Stage Choice: ['A: Not comparable', 'B: The same', 'C: Darker', 'D: Brighter
']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 16.5, 'B': 17.75, 'C': 16.375, 'D': 15.0625}
Dependent Stage Generated Answer: B
--------------------------------
Dependent Stage Name: two_panels_2_3_left
Dependent Stage Question: Consider only the left part of the two panels in the image.
Is the color of all the objects in the left panel the same as, darker or brighter than
the objects in the right panel? If the colors within either panel are already different
 from each other, select 'Not Comparable.'
Dependent Stage Choice: ['A: Not comparable', 'B: Darker', 'C: Brighter', 'D: The same
']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 18.875, 'B': 18.0, 'C': 16.125, 'D': 19.5}
Dependent Stage Generated Answer: D
--------------------------------
Current Stage:
Current Stage Name: one_row_left
Current Stage Question: Look at the three panels in the image from left to right,
paying attention only to the left portions of each panel, and identify the rule that
controls the color of objects.
Current Stage Choice: ['A: The color of objects in the last panel is the sum of the
colors in the previous two panels.', 'B: The color of objects gradually brightens by a
constant amount each time.', 'C: The color of objects gradually darkens by a constant
amount each time.', 'D: The color of objects in the last panel is the difference
between the colors in the previous two panels.']
Current Stage Ground Truth: B
Current Stage Logits: {'A': 20.0, 'B': 21.375, 'C': 20.5, 'D': 19.75}
Current Stage Generated Answer: B
--------------------------------
Accuracy: 1.0
MSEval: 1.066
MSEval Random Baseline: 1.0
```

Listing 2: An instance of high accuracy but low MSEval occurs since the LLM NVLM-D-72B generates a current-stage answer consistent with the ground truth, while earlier dependent stages produce inconsistent results.

```
--------------------------------
Dependent Stage Name: single_panel_1_left
Dependent Stage Question: What is the shape of the object in the left part of the panel
?
Dependent Stage Choice: ['A: circle', 'B: hexagon', 'C: triangle', 'D: square']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 17.125, 'B': 24.5, 'C': 16.0, 'D': 16.0}
Dependent Stage Generated Answer: B
--------------------------------
Dependent Stage Name: single_panel_2_left
Dependent Stage Question: What is the shape of the object in the left part of the panel
?
Dependent Stage Choice: ['A: triangle', 'B: pentagon', 'C: square', 'D: hexagon']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 17.5, 'B': 23.875, 'C': 17.125, 'D': 15.8125}
Dependent Stage Generated Answer: B
--------------------------------
Dependent Stage Name: single_panel_3_left
Dependent Stage Question: What is the shape of the object in the left part of the panel
?
Dependent Stage Choice: ['A: hexagon', 'B: pentagon', 'C: square', 'D: triangle']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 17.375, 'B': 16.875, 'C': 24.125, 'D': 17.375}
Dependent Stage Generated Answer: C
--------------------------------
Dependent Stage Name: two_panels_1_2_left
Dependent Stage Question: Consider only the left part of the two panels in the image.
Is the shape of all the objects in the left panel have the same, more, or fewer edges
compared to the objects in the right panel? If the shapes within either panel are
already different from each other, select 'Not Comparable.' (Note: The edge number
increases in the following order: triangle, square, pentagon, hexagon, circle)
Dependent Stage Choice: ['A: Not comparable', 'B: Fewer', 'C: The same', 'D: More']
Dependent Stage Ground Truth: D
Dependent Stage Logits: {'A': 21.625, 'B': 20.875, 'C': 20.0, 'D': 22.25}
Dependent Stage Generated Answer: D
--------------------------------
Dependent Stage Name: two_panels_2_3_left
Dependent Stage Question: Consider only the left part of the two panels in the image.
Is the shape of all the objects in the left panel have the same, more, or fewer edges
compared to the objects in the right panel? If the shapes within either panel are
already different from each other, select 'Not Comparable.' (Note: The edge number
increases in the following order: triangle, square, pentagon, hexagon, circle)
Dependent Stage Choice: ['A: The same', 'B: Not comparable', 'C: More', 'D: Fewer']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 20.875, 'B': 21.75, 'C': 21.875, 'D': 20.25}
Dependent Stage Generated Answer: C
--------------------------------
Current Stage:
Current Stage Name: one_row_left
Current Stage Question: Look at the three panels in the image from left to right,
paying attention only to the left portions of each panel, and identify the rule that
controls the shape of objects. (Note: The edge number increases in the following order:
 triangle, square, pentagon, hexagon, circle)
Current Stage Choice: ['A: The edge number of shape gradually decreases by 1.', 'B: The
 edge number of shape gradually increases by 1.', 'C: No clear rule is present.', 'D:
The shape remains constant.']
Current Stage Ground Truth: A
Current Stage Logits: {'A': 21.75, 'B': 22.75, 'C': 19.25, 'D': 16.625}
Current Stage Generated Answer: B
--------------------------------
Accuracy: 0.0
MSEval: 2.506839853582758
MSEval Random Baseline: 1.0
```

Listing 3: An instance of low accuracy but high MSEval arises as the LLM NVLM-D-72B generates a current-stage answer inconsistent with the ground truth, despite earlier dependent stages producing consistent results.

```
--------------------------------
Dependent Stage Name: single_panel_1_left
Dependent Stage Question: What is the shape of the object in the left part of the panel
?
Dependent Stage Choice: ['A: circle', 'B: hexagon', 'C: triangle', 'D: square']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 21.125, 'B': 26.375, 'C': 20.375, 'D': 22.375}
Dependent Stage Generated Answer: B
--------------------------------
Dependent Stage Name: single_panel_2_left
Dependent Stage Question: What is the shape of the object in the left part of the panel
?
Dependent Stage Choice: ['A: triangle', 'B: pentagon', 'C: square', 'D: hexagon']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 22.5, 'B': 25.5, 'C': 21.25, 'D': 24.625}
Dependent Stage Generated Answer: B
--------------------------------
Dependent Stage Name: single_panel_3_left
Dependent Stage Question: What is the shape of the object in the left part of the panel
?
Dependent Stage Choice: ['A: hexagon', 'B: pentagon', 'C: square', 'D: triangle']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 22.875, 'B': 21.125, 'C': 27.5, 'D': 23.625}
Dependent Stage Generated Answer: C
--------------------------------
Dependent Stage Name: two_panels_1_2_left
Dependent Stage Question: Consider only the left part of the two panels in the image.
Is the shape of all the objects in the left panel have the same, more, or fewer edges
compared to the objects in the right panel? If the shapes within either panel are
already different from each other, select 'Not Comparable.' (Note: The edge number
increases in the following order: triangle, square, pentagon, hexagon, circle)
Dependent Stage Choice: ['A: Not comparable', 'B: Fewer', 'C: The same', 'D: More']
Dependent Stage Ground Truth: D
Dependent Stage Logits: {'A': 30.5, 'B': 30.25, 'C': 30.75, 'D': 30.875}
Dependent Stage Generated Answer: D
--------------------------------
Dependent Stage Name: two_panels_2_3_left
Dependent Stage Question: Consider only the left part of the two panels in the image.
Is the shape of all the objects in the left panel have the same, more, or fewer edges
compared to the objects in the right panel? If the shapes within either panel are
already different from each other, select 'Not Comparable.' (Note: The edge number
increases in the following order: triangle, square, pentagon, hexagon, circle)
Dependent Stage Choice: ['A: The same', 'B: Not comparable', 'C: More', 'D: Fewer']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 30.375, 'B': 29.75, 'C': 30.375, 'D': 30.25}
Dependent Stage Generated Answer: C
--------------------------------
Current Stage:
Current Stage Name: one_row_left
Current Stage Question: Look at the three panels in the image from left to right,
paying attention only to the left portions of each panel, and identify the rule that
controls the shape of objects. (Note: The edge number increases in the following order:
 triangle, square, pentagon, hexagon, circle)
Current Stage Choice: ['A: The edge number of shape gradually decreases by 1.', 'B: The
 edge number of shape gradually increases by 1.', 'C: No clear rule is present.', 'D:
The shape remains constant.']
Current Stage Ground Truth: A
Current Stage Logits: {'A': 32.75, 'B': 33.0, 'C': 31.25, 'D': 30.5}
Current Stage Generated Answer: B
--------------------------------
Accuracy: 0.0
MSEval: 2.228
MSEval Random Baseline: 1.0
```

Listing 4: An instance of low accuracy but high MSEval arises as the LLM Intern-VL2-8B generates a current-stage answer inconsistent with the ground truth, despite earlier dependent stages producing consistent results.

```
--------------------------------
Dependent Stage Name: single_panel_1_right
Dependent Stage Question: Are all objects in the right part of the panel of the same
size?
Dependent Stage Choice: ['A: No', 'B: Yes', 'C: Only one object']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 31.125, 'B': 31.125, 'C': 29.625}
Dependent Stage Generated Answer: A
--------------------------------
Dependent Stage Name: single_panel_2_right
Dependent Stage Question: Are all objects in the right part of the panel of the same
size?
Dependent Stage Choice: ['A: No', 'B: Yes', 'C: Only one object']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 31.0, 'B': 31.125, 'C': 29.5}
Dependent Stage Generated Answer: B
--------------------------------
Dependent Stage Name: single_panel_3_right
Dependent Stage Question: Are all objects in the right part of the panel of the same
size?
Dependent Stage Choice: ['A: Only one object', 'B: No', 'C: Yes']
Dependent Stage Ground Truth: A
Dependent Stage Logits: {'A': 31.375, 'B': 31.375, 'C': 31.625}
Dependent Stage Generated Answer: C
--------------------------------
Dependent Stage Name: two_panels_1_2_right
Dependent Stage Question: Consider only the right part of the two panels in the image.
Is the size of all the objects in the left panel the same as, smaller or larger than
the objects in the right panel? If the sizes within either panel are already different
from each other, select 'Not Comparable.
Dependent Stage Choice: ['A: Not comparable', 'B: Smaller', 'C: Larger', 'D: The same']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 30.875, 'B': 30.75, 'C': 30.25, 'D': 30.75}
Dependent Stage Generated Answer: A
--------------------------------
Dependent Stage Name: two_panels_2_3_right
Dependent Stage Question: Consider only the right part of the two panels in the image.
Is the size of all the objects in the left panel the same as, smaller or larger than
the objects in the right panel? If the sizes within either panel are already different
from each other, select 'Not Comparable.
Dependent Stage Choice: ['A: Not comparable', 'B: Smaller', 'C: The same', 'D: Larger']
Dependent Stage Ground Truth: D
Dependent Stage Logits: {'A': 30.25, 'B': 30.125, 'C': 30.5, 'D': 30.125}
Dependent Stage Generated Answer: C
--------------------------------
Current Stage:
Current Stage Name: one_row_right
Current Stage Question: Analyze the three panels in the image from left to right,
concentrating only on the right areas of each panel, and determine the rule that
dictates the size of objects.
Current Stage Choice: ['A: The size of objects gradually decreases by a constant amount
 each time.', 'B: The size of objects in the last panel is the difference between the
sizes in the previous two panels.', 'C: The size remains constant.', 'D: Three distinct
 sizes across panels, rotating through each possible permutation of these sizes.']
Current Stage Ground Truth: D
Current Stage Logits: {'A': 31.125, 'B': 30.375, 'C': 31.0, 'D': 31.375}
Current Stage Generated Answer: D
--------------------------------
Accuracy: 1.0
MSEval: 0.955
MSEval Random Baseline: 1.0
```

Listing 5: An instance of high accuracy but low MSEval occurs since the LLM Intern-VL2-8B generates a current-stage answer consistent with the ground truth, while earlier dependent stages produce inconsistent results.

```
--------------------------------
Dependent Stage Name: single_panel_1_right
Dependent Stage Question: What is the shape of the object in the right part of the
panel?
Dependent Stage Choice: ['A: circle', 'B: square', 'C: hexagon', 'D: pentagon']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 0, 'B': 0, 'C': 30.625, 'D': 0}
Dependent Stage Generated Answer: C
--------------------------------
Dependent Stage Name: single_panel_2_right
Dependent Stage Question: What is the shape of the object in the right part of the
panel?
Dependent Stage Choice: ['A: pentagon', 'B: triangle', 'C: square', 'D: circle']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 0, 'B': 30.5, 'C': 0, 'D': 0}
Dependent Stage Generated Answer: B
--------------------------------
Dependent Stage Name: single_panel_3_right
Dependent Stage Question: What is the shape of the object in the right part of the
panel?
Dependent Stage Choice: ['A: triangle', 'B: hexagon', 'C: square', 'D: circle']
Dependent Stage Ground Truth: C
Dependent Stage Logits: {'A': 0, 'B': 0, 'C': 30.875, 'D': 0}
Dependent Stage Generated Answer: C
--------------------------------
Dependent Stage Name: two_panels_1_2_right
Dependent Stage Question: Consider only the right part of the two panels in the image.
Is the shape of all the objects in the left panel have the same, more, or fewer edges
compared to the objects in the right panel? If the shapes within either panel are
already different from each other, select 'Not Comparable.' (Note: The edge number
increases in the following order: triangle, square, pentagon, hexagon, circle)
Dependent Stage Choice: ['A: The same', 'B: More', 'C: Fewer', 'D: Not comparable']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 0, 'B': 30.0, 'C': 0, 'D': 0}
Dependent Stage Generated Answer: B
--------------------------------
Dependent Stage Name: two_panels_2_3_right
Dependent Stage Question: Consider only the right part of the two panels in the image.
Is the shape of all the objects in the left panel have the same, more, or fewer edges
compared to the objects in the right panel? If the shapes within either panel are
already different from each other, select 'Not Comparable.' (Note: The edge number
increases in the following order: triangle, square, pentagon, hexagon, circle)
Dependent Stage Choice: ['A: Not comparable', 'B: Fewer', 'C: More', 'D: The same']
Dependent Stage Ground Truth: B
Dependent Stage Logits: {'A': 0, 'B': 28.25, 'C': 0, 'D': 0}
Dependent Stage Generated Answer: B
--------------------------------
Current Stage:
Current Stage Name: one_row_right
Current Stage Question: Inspect the three panels in the image from left to right,
focusing exclusively on the right parts of each panel, and uncover the rule that
governs the shape of objects. (Note: The edge number increases in the following order:
triangle, square, pentagon, hexagon, circle)
Current Stage Choice: ['A: The shape remains constant.', 'B: Three distinct shapes
across panels, rotating through each possible permutation of these shapes.', 'C: No
clear rule is present.', 'D: The edge number of shape gradually increases by 1.']
Current Stage Ground Truth: B
Current Stage Logits: {'A': 0, 'B': 0, 'C': 31.67, 'D': 0}
Current Stage Generated Answer: C
--------------------------------
Accuracy: 0.0
MSEval: 2.345087186311839
MSEval Random Baseline: 1.0
```

Listing 6: An instance of low accuracy but high MSEval arises as the LLM Qwen2-VL-72B generates a current-stage answer inconsistent with the ground truth, despite earlier dependent stages producing consistent results.