VISULOGIC: A BENCHMARK FOR EVALUATING VISUAL REASONING IN MULTI-MODAL LARGE LANGUAGE MODELS

Weiye Xu[1,3*†], **Jiahao Wang**[2,3*†], **Weiyun Wang**[3†], **Zhe Chen**[3†], **Wengang Zhou**[1], **Aijun Yang**[2], **Lewei Lu**[4], **Houqiang Li**[1], **Xiaohua Wang**[2], **Xizhou Zhu**[3], **Wenhai Wang**[3], **Jifeng Dai**[5,3✉], **Jinguo Zhu**[3✉]
[1]University of Science and Technology of China, [2]Xi'an Jiaotong University,
[3]Shanghai Artifcial Intelligence Laboratory, [4]SenseTime Research, [5]Tsinghua University

ABSTRACT

Visual reasoning is a core component of human intelligence and a critical capability for advanced multimodal models. Yet current reasoning evaluations of multimodal large language models (MLLMs) often rely on text descriptions and allow language-based reasoning shortcuts, failing to measure genuine vision-centric reasoning. To address this, we introduce VisuLogic: a benchmark of 1,000 human-verified problems across six categories (e.g., quantitative shifts, spatial relations, attribute comparisons). These various types of questions can be evaluated to assess the visual reasoning capabilities of MLLMs from multiple perspectives. We evaluate leading MLLMs on this benchmark and analyze their results to identify common failure modes. Most models score below 30% accuracy—only slightly above the 25% random baseline and far below the 51.4% achieved by humans—revealing significant gaps in visual reasoning. Furthermore, we provide a supplementary training dataset and a reinforcement-learning baseline to support further progress. Code, data, and baselines are available at https://visulogic-benchmark.github.io/VisuLogic.

1 INTRODUCTION

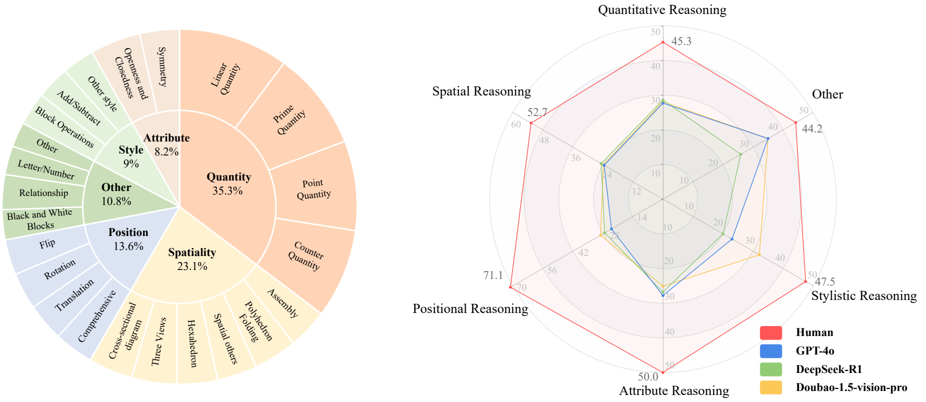
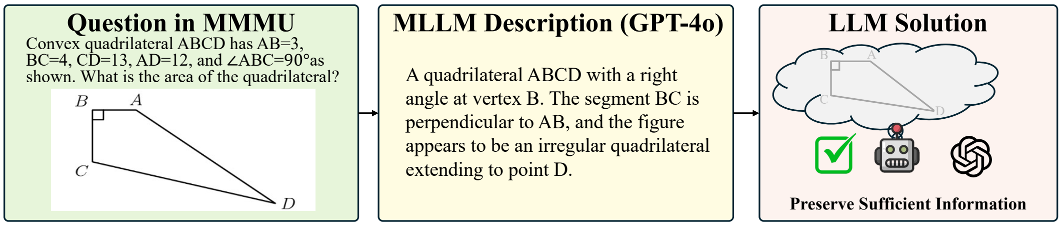


Figure 1: **Composition of the VisuLogic benchmark and performance of representative MLLMs.** The left figure shows the distribution of the 6 categories and their subcategories in VisuLogic. The right figure shows accuracies (%) achieved by MLLMs and by human on each category of VisuLogic.

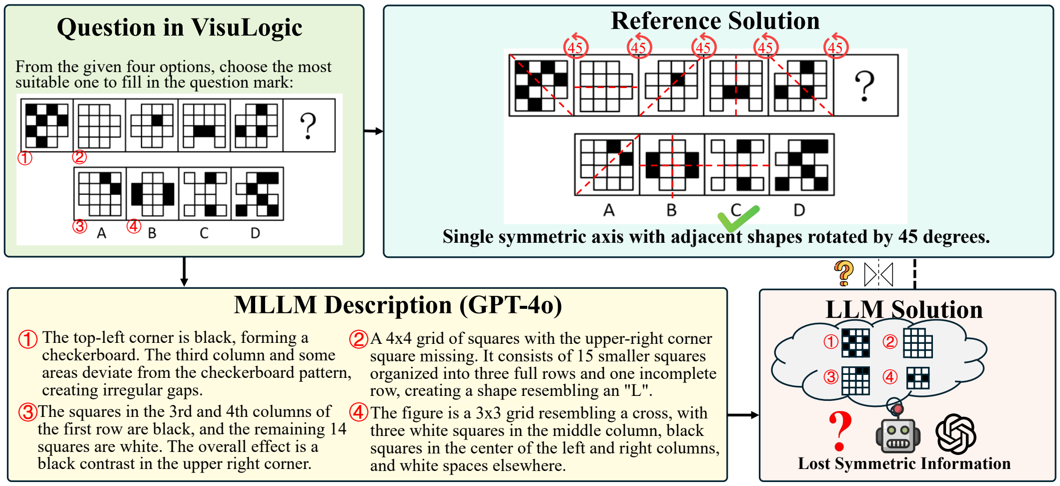
Reasoning, as fundamental component of human intelligence, has become a critical criterion in evaluating progress toward Artificial General Intelligence (AGI) [28, 78]. Recent advancements in Large Language Models (LLMs) have demonstrated substantial improvements in reasoning capabilities across complex domains such as mathematics [64, 86, 85, 61], logical reasoning [72, 83,



Convex quadrilateral ABCD has AB=3, BC=4, CD=13, AD=12, and ∠ABC=90°as shown. What is the area of the quadrilateral?

A quadrilateral ABCD with a right angle at vertex B. The segment BC is perpendicular to AB, and the figure appears to be an irregular quadrilateral extending to point D.

(a) Pipeline of "MLLM description→LLM" for Question in MMMU [93]. It is trivial that SOTA MLLMs extract key visual details, thereby enabling the LLM to answer questions solely based on language reasoning.


Question in VisuLogic
①
②
?
③ A
④ B
C
D
Reference Solution
45
45
45
45
45
?
A
B
C
D
Single symmetric axis with adjacent shapes rotated by 45 degrees.
MLLM Description (GPT-4o)
LLM Solution
①
②
③
④
?
Lost Symmetric Information

From the given four options, choose the most suitable one to fill in the question mark:

① The top-left corner is black, forming a checkerboard. The third column and some areas deviate from the checkerboard pattern, creating irregular gaps.

② A 4x4 grid of squares with the upper-right corner square missing. It consists of 15 smaller squares organized into three full rows and one incomplete row, creating a shape resembling an "L".

③ The squares in the 3rd and 4th columns of the first row are black, and the remaining 14 squares are white. The overall effect is a black contrast in the upper right corner.

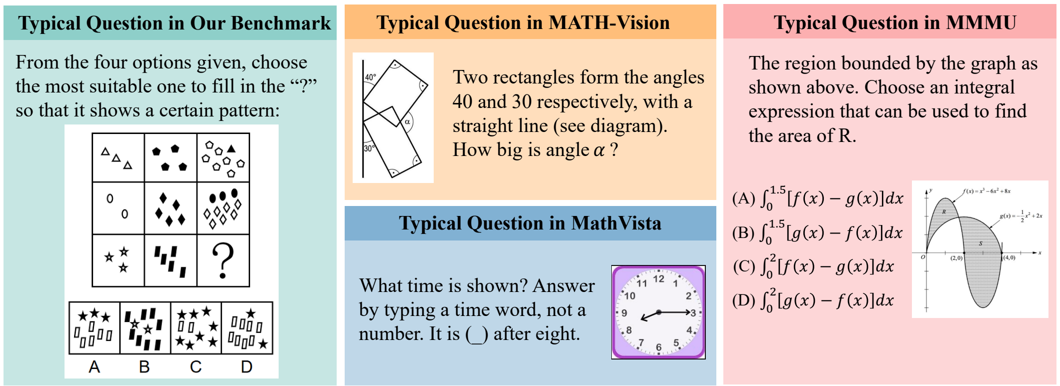
④ The figure is a 3x3 grid resembling a cross, with three white squares in the middle column, black squares in the center of the left and right columns, and white spaces elsewhere.

(b) Pipeline of "MLLM description→LLM" for Question in VisuLogic. Even SOTA MLLMs struggle to describe images precisely, leading to ambiguous interpretations.

Figure 2: **Comparison of the "MLLM description→LLM" pipeline on two benchmarks.** In MMMU, detailed descriptions lead to correct solutions, while in VisuLogic, critical visual cues (*e.g.*, symmetry, rotation) can be easily lost, causing the LLM to misinterpret the image. This highlights that textual reasoning alone is insufficient, underscoring the benchmark's demand for robust and in-depth visual reasoning.

25, 50] and coding [2, 37, 44, 34]. Techniques like Chain-of-Thought (CoT) [79] prompting and test-time compute scaling (e.g., OpenAI o1 [36] and Deepseek-R1 [20]) have significantly enhanced the reasoning performance of LLMs [20, 28, 78]. Along with the rapid development of language reasoning research for LLMs, considerable progress [88, 64, 61, 13, 53, 76, 54, 66, 77, 6, 47] has been made in improving multimodal reasoning capability of Multimodal Large Language Models (MLLMs).

These methods, which often incorporate reinforcement learning techniques [13, 53, 64] to enhance the reasoning capabilities of MLLMs, have achieved some early successes [88, 64, 61, 13, 53, 54, 66]. However, they typically rely on existing multi-modal benchmarks that struggle to accurately capture a model's core visual reasoning ability. For example, VLM-R1 [66] assesses "visual reasoning" with referring expression comprehension tasks [92, 58, 40], yet these tasks primarily focus on object localization, demanding only basic perceptual skills rather than more advanced visual cognitive processes. Meanwhile, several works [61, 64, 88] adopt mathematical problem-solving benchmarks that include diagrams—such as MathVista [55], MathVerse [95], and MathVision [73]—to evaluate visual reasoning. In practice, however, as [95] observes, many MLLMs translate these visual clues into textual descriptions and then rely on standard language reasoning. This approach can incorrectly attribute language-driven results to visual reasoning, resulting in a misleading assessment of the model's visual reasoning capabilities [95, 32]. Consequently, designing new benchmarks that explicitly focus on vision-centric reasoning—rather than conflating it with text-based reasoning—remains critical for advancing MLLMs' visual reasoning capacities.



Typical Question in Our Benchmark

From the four options given, choose the most suitable one to fill in the "?" so that it shows a certain pattern:

Typical Question in MATH-Vision

Two rectangles form the angles 40 and 30 respectively, with a straight line (see diagram). How big is angle α ?

Typical Question in MathVista

What time is shown? Answer by typing a time word, not a number. It is (_) after eight.

Typical Question in MMMU

The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

(A) $\int_0^{1.5}[f(x) - g(x)]dx$
(B) $\int_0^{1.5}[g(x) - f(x)]dx$
(C) $\int_0^{2}[f(x) - g(x)]dx$
(D) $\int_0^{2}[g(x) - f(x)]dx$

Figure 3: **Comparison of questions from different Benchmarks.** Compared to MathVista [55], MathVision [73], and MMMU [93], VisuLogic focuses more explicitly on assessing pure visual reasoning capabilities.

To address this limitation, we propose VisuLogic, a novel benchmark specifically designed to evaluate visual reasoning abilities in multimodal models without mixing them with purely text-based reasoning (see Figure 3). VisuLogic comprises carefully constructed tasks that span multiple reasoning categories (see Figure 1). As shown in Figure 5, these tasks are classified into six key types, such as Quantitative Reasoning, which requires understanding and deducing shifts in the quantity of certain elements within an image. In contrast to existing benchmarks, as demonstrated in Figure 2, state-of-the-art (SOTA) MLLMs often omit crucial visual details when describing VisuLogic problems, making it difficult for them to rely solely on a text-based inference shortcut. Indeed, even humans would find it challenging to capture every essential visual cue in a single description, so effectively tackling VisuLogic demands more robust, vision-centric reasoning. By reducing reliance on textual inference shortcuts, VisuLogic thus provides a stringent evaluation of MLLMs' genuine visual reasoning capabilities.

We conducted a comprehensive evaluation and systematic analysis to assess current models' visual reasoning capabilities. When leading text-only LLMs were supplied with detailed descriptions in place of raw images, their accuracy—Doubao-1.5-Pro (26.6%), Claude-3.7-Sonnet (25.9%) and Qwen2.5-72B-Instruct [87] (28.0%)—barely exceeded the random-chance baseline of 24.9%. This clearly demonstrates that textual reasoning alone are insufficient for solving our VisuLogic tasks. Even state-of-the-art multimodal large language models (MLLMs)—including OpenAI-o3, GPT-4o [35], Gemini-2.0-Pro-Exp [68] and InternVL3-78B [98]—achieve only 29.5%, 26.3%, 28.0% and 27.7%, respectively, whereas human participants reached 51.4%. The substantial gap between these results and human performance underscores the challenge of robust visual reasoning in current MLLMs. To probe the limits of these models further, we ran "hint" experiments in which explicit problem-solving cues are provided. Under such conditions, human accuracy rose to 83.6%, yet MLLMs still failed to surpass 50.0%.

2 RELATED WORK

Multi-modal Large Language Models. Recent years have witnessed substantial advancements in Multi-modal Large Language Models (MLLMs). Early works like BLIP [43, 42] and Flamingo [5] introduce lightweight parameters between vision transformer [23] (ViT) and LLMs, laying the groundwork for multimodal perception. Subsequent efforts, such as LLaVA [48] and MiniGPT-4 [97], integrate instruction tuning, further enhancing the performance of MLLMs. Proprietary models like GPT-4o [35] and Gemini-Pro [68] have advanced MLLM performance on complex multimodal tasks, while open-source models such as Qwen-VL series [8, 74, 9] and InternVL series [17, 18, 26, 16, 98] achieve competitive results through optimized architectural design, dataset expansion and training paradigm improvements. Meanwhile, some related studies further advance the ability of large models by incorporating new modalities (e.g., audio [24, 21, 81], point clouds [29, 11], video [96, 14]) and by supporting more tasks (e.g., grounding [84, 75], computer usage [63, 7]). Notably, limited research attempts to enhance the reasoning capabilities of MLLMs. Some pioneering works, such as

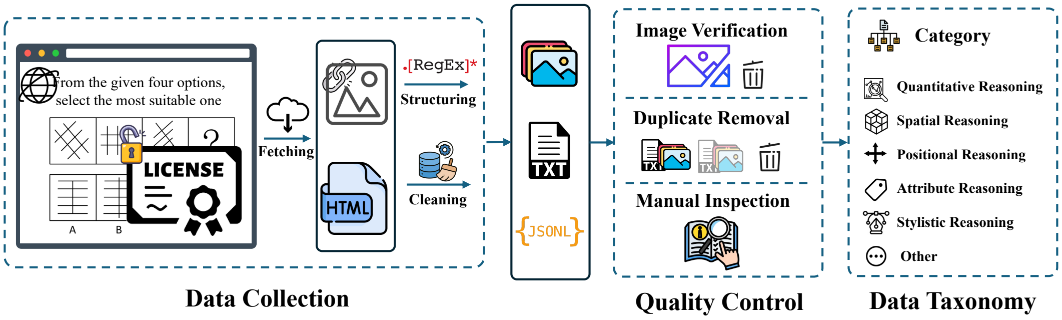


Figure 4: **Data curation pipeline of VisuLogic.** The pipeline includes Data Collection, Quality Control and Data Taxonomy.

R1-Onevision [88], LMM-R1 [64], MM-EUREKA [61], R1-V [13], Visual-rft [53], Visualprm [76], OThink-MR1 [54], VLM-R1 [66], and Open-r1-Video [77] have explored the visual reasoning capabilities of MLLMs through Reinforcement Learning (RL), but they are still in the nascent stage.

Multimodal Benchmarks. With the development of MLLMs, multimodal benchmarks have also evolved significantly [45]. Early benchmarks primarily address visual perception tasks through simple tasks like visual question answering (VQA) [15, 46, 38, 82], image captioning [62, 22, 39] and referring expression comprehension [92, 58]. Subsequent works expand the capability coverage of benchmarks into more specialized domains: OCRBench [52], Chartqa [59] and DocVQA [60] assess textual content extraction; AgentBench [51] and ToolEyes [90] test tool usage capabilities; and egocentric perception benchmarks [57, 19] quantify first-person scene interpretation. Despite the progress, they ignore the evaluation of visual reasoning abilities [94, 93]. Recently, some benchmarks have made explorations in examining MLLMs' visual reasoning abilities, but methodological deficiencies still cause limitations to assess the intrinsic visual reasoning capabilities [32, 4, 80]. InfiMM-Eval [31] test reasoning abilities around daily life, lacking deep-level reasoning scenarios. MMMU [93] and Emma [32] provide benchmarks demanding advanced reasoning abilities in fields such as chemistry and physics, but they ignore questions around the images' fundamental visual components (e.g., shapes, elements). While mathematical benchmarks [73, 55, 33, 65, 95, 30] evaluate mathematical reasoning with geometric and diagram problems included, they focus on math capabilities but disregard logical analysis about the vision information. LogicVista [80] provides a multimodal logical reasoning benchmark, its visual questions lack analytical depth—dominated by single-hop, superficial queries in limited data scope. Unlike previous works, we introduce a challenging benchmark focused specifically on the domain of visual logical reasoning.

3 VISULOGIC

In this section, we first describe the VisuLogic data-curation pipeline, which comprises three key stages: data collection, quality control, and the detailed taxonomy. We then report the benchmark's construction statistics, including total size, answer-option distributions, and category-level proportions. Finally, we introduce a supplementary training dataset—consisting of questions analogous to those in VisuLogic—designed to bolster future research and facilitate community engagement.

3.1 DATA CURATION PIPELINE

Data Collection. We construct the VisuLogic dataset by sourcing all questions from publicly available online resources in compliance with relevant licenses and regulations. As shown in Figure 4, our automated data processing pipeline comprises three stages: 1) **Fetching**: We employ Playwright to systematically scrape raw web content, supplemented by custom parsing scripts that extract question–answer pairs. 2) **Cleaning**: We remove noise, irrelevant content, and extraneous HTML markup (*e.g.*, `<div>`) to ensure the integrity of the textual data. 3) **Structuring**: We standardize the cleaned text and images by structuring all information JSONL format.

Quality Control. To ensure the reliability of the benchmark dataset, we employ a three-stage data validation procedure: 1) **Image Verification**: Each image referenced in the questions is checked for existence and correct formatting; any item that fails to meet the criteria is removed following human review. 2) **Duplicate Removal**: We eliminate redundant entries at both the text and image levels by (i) detecting lexical overlap among text segments and (ii) applying perceptual hashing (pHash) to identify visually similar images. 3) **Manual Checking**: After automated filtering, we perform a thorough human-led review of every remaining entry to confirm its validity and ensure reliability.

Data Taxonomy. We categorize all collected data into a taxonomy of six primary classes based on expert human annotation of the reasoning skills each question requires. Annotators first tag questions according to the targeted reasoning competency; these annotated tags are then analyzed and merged into five primary categories. A subsequent human review ensures that every question is accurately classified, with any ambiguous instances consolidated under the "Other" category. Specifically, we define each category as follows. **Quantitative Reasoning** focuses on changes in the number or count of graphical elements (for example, points, lines and angles) and on arithmetic relationships among shapes. **Spatial Reasoning** requires mentally reconstructing three-dimensional shapes from two-dimensional figures, folding or unfolding surfaces, and integrating three-dimensional structures. **Positional Reasoning** examines transformations such as translation, rotation and reflection of objects while preserving their fundamental elements. **Attribute Reasoning** involves intrinsic properties of shapes, including symmetry (axial or central), curvature and measures of openness or closedness. **Stylistic Reasoning** entails alterations in stylistic features such as overlay, subtraction and assessments of shape similarity or difference. **Other** includes questions that fall outside the preceding categories, including those involving letters, alphanumeric symbols or specialized characters.

3.2 BENCHMARK DATASET STATISTICS

The VisuLogic benchmark comprises 1,000 rigorously validated single-choice visual-reasoning questions spanning six categories—Quantitative (35.3%), Spatial (23.1%), Positional (13.6%), Attribute (8.2%), Stylistic (9.0%), and Other (10.8%) with correct answers evenly balanced across options ABCD (23.1%, 26.7%, 25.2%, 25.0%).

3.3 SUPPLEMENTARY TRAINING DATASET

To facilitate further investigation of visual reasoning, we provide an auxiliary training set of 4,296 question–answer pairs drawn from the same domains and subjected to identical validation procedures to prevent overlap with the benchmark. The training split mirrors the primary taxonomy, with category proportions of Quantitative Reasoning (30.7%), Spatial Reasoning (25.5%), Positional Reasoning (13.0%), Attribute Reasoning (8.8%), Stylistic Reasoning (9.9%), and Other (12.1%).

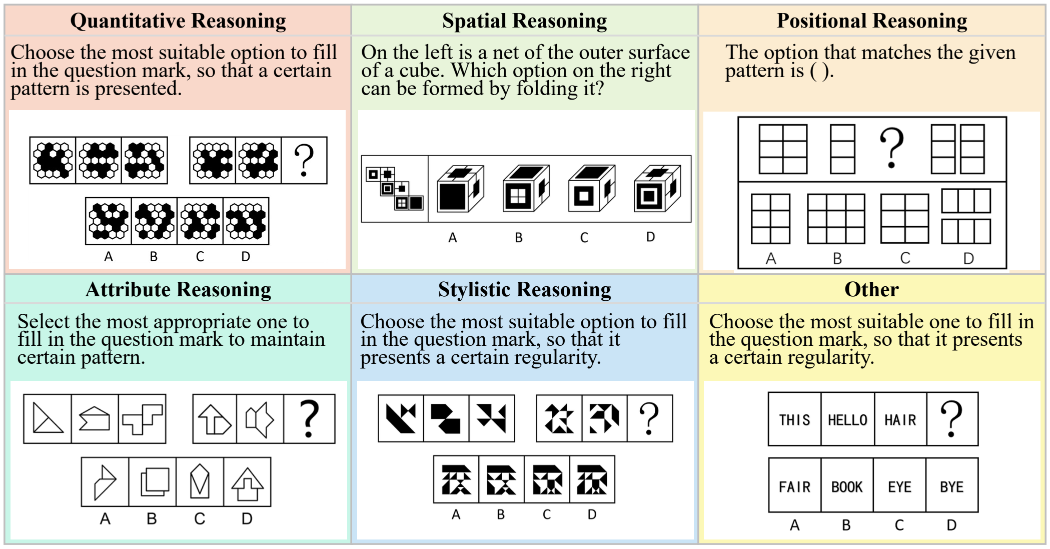
4 EXPERIMENTS

In this section, we present a comprehensive evaluation of the VisuLogic benchmark. We first describe the experimental setup in Section 4.1, followed by overall performance results in Section 4.2. We then analyze systematic errors in Section 4.3 and provide qualitative insights in Section 4.4.

4.1 EXPERIMENT SETUP

References Performance. To fully investigate models' performance, we establish two reference points: 1) **Human Performance**: We invited 100 graduate students majoring in science and engineering to solve 10 randomly sampled VisuLogic questions each, allowing 2–5 minutes per question. The aggregate accuracy over all participants constitutes the human benchmark. 2) **Random Selection**: We simulate random guessing by sampling answers uniformly over 10 independent runs and report the average accuracy as the random baseline.

Evaluated Models. We evaluate a total of 31 models on VisuLogic, comprising 8 large language models (LLMs) and 23 multimodal large language models (MLLMs). Appendix C provides more details.



Quantitative Reasoning

Choose the most suitable option to fill in the question mark, so that a certain pattern is presented.

Spatial Reasoning

On the left is a net of the outer surface of a cube. Which option on the right can be formed by folding it?

Positional Reasoning

The option that matches the given pattern is ( ).

Attribute Reasoning

Select the most appropriate one to fill in the question mark to maintain certain pattern.

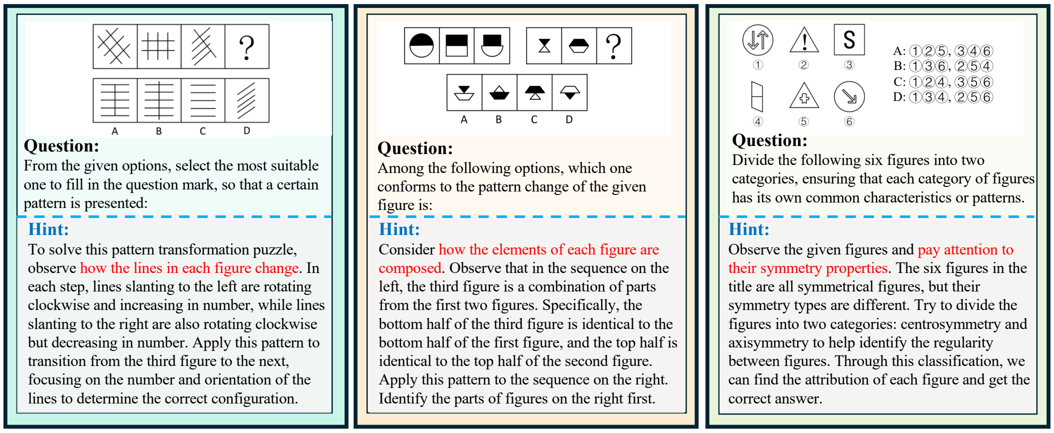
Stylistic Reasoning

Choose the most suitable option to fill in the question mark, so that it presents a certain regularity.

Other

Choose the most suitable one to fill in the question mark, so that it presents a certain regularity.

Figure 5: **Question examples of different categories in our VisuLogic Benchmark.** VisuLogic contains 6 categories of questions, which require models' abilities in visual logic reasoning.



Question:
From the given options, select the most suitable one to fill in the question mark, so that a certain pattern is presented:

Hint:
To solve this pattern transformation puzzle, observe how the lines in each figure change. In each step, lines slanting to the left are rotating clockwise and increasing in number, while lines slanting to the right are also rotating clockwise but decreasing in number. Apply this pattern to transition from the third figure to the next, focusing on the number and orientation of the lines to determine the correct configuration.

Question:
Among the following options, which one conforms to the pattern change of the given figure is:

Hint:
Consider how the elements of each figure are composed. Observe that in the sequence on the left, the third figure is a combination of parts from the first two figures. Specifically, the bottom half of the third figure is identical to the bottom half of the first figure, and the top half is identical to the top half of the second figure. Apply this pattern to the sequence on the right. Identify the parts of figures on the right first.

A: ①②⑤, ③④⑥
B: ①③⑥, ②⑤④
C: ①②④, ③⑤⑥
D: ①③④, ②⑤⑥

Question:
Divide the following six figures into two categories, ensuring that each category of figures has its own common characteristics or patterns.

Hint:
Observe the given figures and pay attention to their symmetry properties. The six figures in the title are all symmetrical figures, but their symmetry types are different. Try to divide the figures into two categories: centrosymmetry and axisymmetry to help identify the regularity between figures. Through this classification, we can find the attribution of each figure and get the correct answer.

Figure 6: **Hint prompts visualization.** Hint prompts examples, which supply solution guidance for MLLMs, are shown in the image, with solution-critical elements highlighted in red.

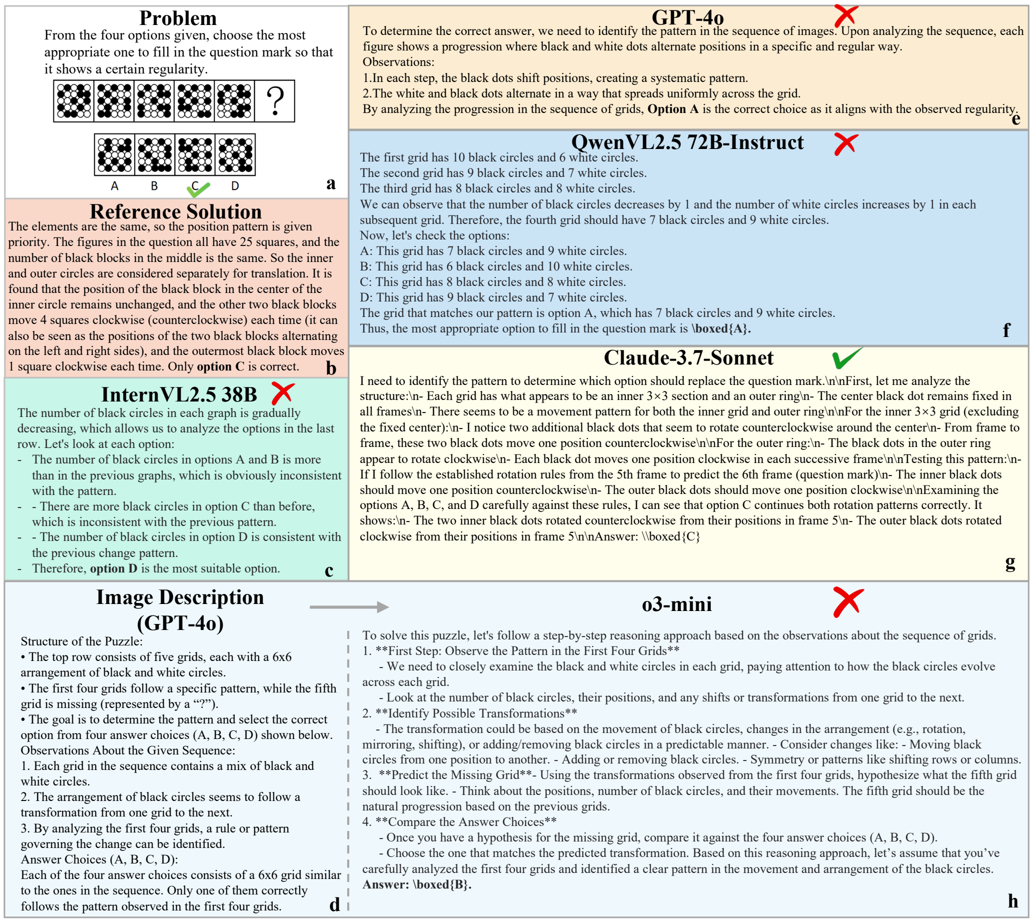
LLM Evaluation Protocol. For language models, we generate an image description using GPT-4o and prepend it to the question. Specifically, each question is formatted as "*Following is a detailed caption describing an image: [DESCRIPTION]. Based on the provided description, select the best answer from the four options.*". This combined prompt is fed directly into LLMs for inference.

Prompts Setting. We apply three distinct prompting paradigms to investigate model reasoning capabilities:

1) **Non-CoT prompt evaluation**: Models receive a concise instruction: "*Answer the question using a single word or phrase, following this format: Answer: \boxed{$LETTER}*".

2) **CoT prompt evaluation**: We prompt models to articulate intermediate reasoning steps: "*Solve the complex visual logical reasoning problem through step-by-step reasoning. Think about the reasoning process first and answer the question following this format: Answer: \boxed{$LETTER}*".

3) **Hint prompts evaluation**: Leveraging GPT-4o, we generate question-specific hints derived from the reference solutions. Solution-related hints are provided alongside the CoT prompt to guide reasoning without revealing the final answer directly.



Problem

From the four options given, choose the most appropriate one to fill in the question mark so that it shows a certain regularity.

Reference Solution

The elements are the same, so the position pattern is given priority. The figures in the question all have 25 squares, and the number of black blocks in the middle is the same. So the inner and outer circles are considered separately for translation. It is found that the position of the black block in the center of the inner circle remains unchanged, and the other two black blocks move 4 squares clockwise (counterclockwise) each time (it can also be seen as the positions of the two black blocks alternating on the left and right sides), and the outermost black block moves 1 square clockwise each time. Only **option C** is correct.

InternVL2.5 38B ✗

The number of black circles in each graph is gradually decreasing, which allows us to analyze the options in the last row. Let's look at each option:

- The number of black circles in options A and B is more than in the previous graphs, which is obviously inconsistent with the pattern.
- - There are more black circles in option C than before, which is inconsistent with the previous pattern.
- - The number of black circles in option D is consistent with the previous change pattern.
- Therefore, **option D** is the most suitable option.

Image Description (GPT-4o)

Structure of the Puzzle:
• The top row consists of five grids, each with a 6x6 arrangement of black and white circles.
• The first four grids follow a specific pattern, while the fifth grid is missing (represented by a "?").
• The goal is to determine the pattern and select the correct option from four answer choices (A, B, C, D) shown below.
Observations About the Given Sequence:
1. Each grid in the sequence contains a mix of black and white circles.
2. The arrangement of black circles seems to follow a transformation from one grid to the next.
3. By analyzing the first four grids, a rule or pattern governing the change can be identified.
Answer Choices (A, B, C, D):
Each of the four answer choices consists of a 6x6 grid similar to the ones in the sequence. Only one of them correctly follows the pattern observed in the first four grids.

GPT-4o ✗

To determine the correct answer, we need to identify the pattern in the sequence of images. Upon analyzing the sequence, each figure shows a progression where black and white dots alternate positions in a specific and regular way.
Observations:
1.In each step, the black dots shift positions, creating a systematic pattern.
2.The white and black dots alternate in a way that spreads uniformly across the grid.
By analyzing the progression in the sequence of grids, **Option A** is the correct choice as it aligns with the observed regularity.

QwenVL2.5 72B-Instruct ✗

The first grid has 10 black circles and 6 white circles.
The second grid has 9 black circles and 7 white circles.
The third grid has 8 black circles and 8 white circles.
We can observe that the number of black circles decreases by 1 and the number of white circles increases by 1 in each subsequent grid. Therefore, the fourth grid should have 7 black circles and 9 white circles.
Now, let's check the options:
A: This grid has 7 black circles and 9 white circles.
B: This grid has 6 black circles and 10 white circles.
C: This grid has 8 black circles and 8 white circles.
D: This grid has 9 black circles and 7 white circles.
The grid that matches our pattern is option A, which has 7 black circles and 9 white circles.
Thus, the most appropriate option to fill in the question mark is **\boxed{A}.**

Claude-3.7-Sonnet ✓

I need to identify the pattern to determine which option should replace the question mark.\n\nFirst, let me analyze the structure:\n- Each grid has what appears to be an inner 3×3 section and an outer ring\n- The center black dot remains fixed in all frames\n- There seems to be a movement pattern for both the inner grid and outer ring\n\nFor the inner 3×3 grid (excluding the fixed center):\n- I notice two additional black dots that seem to rotate counterclockwise around the center\n- From frame to frame, these two black dots move one position counterclockwise\n\nFor the outer ring:\n- The black dots in the outer ring appear to rotate clockwise\n- Each black dot moves one position clockwise in each successive frame\n\nTesting this pattern:\n- If I follow the established rotation rules from the 5th frame to predict the 6th frame (question mark)\n- The inner black dots should move one position counterclockwise\n- The outer black dots should move one position clockwise\n\nExamining the options A, B, C, and D carefully against these rules, I can see that option C continues both rotation patterns correctly. It shows:\n- The two inner black dots rotated counterclockwise from their positions in frame 5\n- The outer black dots rotated clockwise from their positions in frame 5\n\nAnswer: \\boxed{C}

o3-mini ✗

To solve this puzzle, let's follow a step-by-step reasoning approach based on the observations about the sequence of grids.
1. **First Step: Observe the Pattern in the First Four Grids**
 - We need to closely examine the black and white circles in each grid, paying attention to how the black circles evolve across each grid.
 - Look at the number of black circles, their positions, and any shifts or transformations from one grid to the next.
2. **Identify Possible Transformations**
 - The transformation could be based on the movement of black circles, changes in the arrangement (e.g., rotation, mirroring, shifting), or adding/removing black circles in a predictable manner. - Consider changes like: - Moving black circles from one position to another. - Adding or removing black circles. - Symmetry or patterns like shifting rows or columns.
3. **Predict the Missing Grid**- Using the transformations observed from the first four grids, hypothesize what the fifth grid should look like. - Think about the positions, number of black circles, and their movements. The fifth grid should be the natural progression based on the previous grids.
4. **Compare the Answer Choices**
 - Once you have a hypothesis for the missing grid, compare it against the four answer choices (A, B, C, D).
 - Choose the one that matches the predicted transformation. Based on this reasoning approach, let's assume that you've carefully analyzed the first four grids and identified a clear pattern in the movement and arrangement of the black circles.
Answer: \boxed{B}.

Figure 7: **Solution examples generated by different models.** Reference solution and outputs generated by GPT-4o [35], Qwen2.5VL-72B-Instruct [9], InternVL2.5-38B [18], and Claude-3.7-Sonnet. Additionally, the image description and solution from LLMs are also illustrated.

Notably, unless otherwise specified, CoT prompt evaluation is employed by default for assessing model performance. And hint prompt cases can be found in Figure 6.

4.2 OVERALL RESULTS

LLM Performance. Table 1 reports that all evaluated LLMs attain rather low accuracy on VisuLogic. The best-performing LLM, *Qwen2.5-72B-Instruct*, reaches only 28.0%, while *GPT-4* and *Deepseek-R1* achieve 23.6% and 26.6%, respectively. These findings underscore that reasoning based solely on textual descriptions is insufficient to capture the rich visual information required by our benchmark, causing failures to resolve visual logical reasoning problems.

MLLM Performance. As shown in Table 1, current multimodal LLMs also perform poorly on VisuLogic. The highest score is 29.5% by *OpenAI-o3*, which remains a substantial 21.9 points below human performance. Advanced models such as *GPT-4o* and *Gemini-2.0-Pro* attain only 26.3% and 28.0%, respectively, revealing a marked gap between existing MLLMs and human-level visual reasoning. Overall, these results indicate that current MLLMs have serious deficiencies in visual reasoning and that significant advances are still required.

Effectiveness of CoT Prompts. Contrary to expectations, chain-of-thought (CoT) prompting yields minimal improvements in visual reasoning. As detailed in Table 2, *GPT-4o-mini* benefits most, with only a 1.2-point gain under CoT compared to direct-answer prompts; all other models exhibit gains below 1.0 point. We speculate that this limited effect likely stems from current CoT training being

Table 1: **Cross-Modal performance with CoT prompts on VisuLogic.** The table shows the evaluation scores of baseline references, LLMs, and MLLMs. Top performers per category are **bolded**, second - place ones underlined. The first column shows the model name. The second shows the total score of the Hint prompt with details in Table 8. The third shows the total score with the CoT prompt. And the remaining columns show the category-wise scores with the CoT prompt.

Models	Hint	Overall	Quantity	Spatiality	Position	Attribute	Style	Other
References								
Human	83.6	51.4	45.3	52.7	71.1	50.0	47.5	44.2
Random	25.0	24.9	25.7	25.4	22.7	23.4	24.3	26.1
Open Source LLM (MLLM Description→LLM)								
Deepseek-R1 [20]	-	26.6	27.7	23.5	24.0	**27.8**	23.0	**35.0**
Qwen2.5-72B-Instruct [87]	-	**28.0**	**30.2**	**24.4**	**27.5**	26.5	**26.8**	30.8
QwQ-32B [71]	-	22.8	24.6	20.1	25.4	19.0	20.7	24.0
Close Source LLM (MLLM Description→LLM)								
GPT-4 (20240613) [1]	-	23.6	21.2	22.5	21.3	25.6	23.3	**35.2**
o3-mini (20250131)	-	24.6	27.8	18.8	24.5	21.7	25.6	28.4
Gemini-2.0-Flash-Thinking (20250121) [68]	-	23.4	23.2	**26.0**	16.9	17.1	21.1	33.3
Claude-3.7-Sonnet (20250219)	-	25.9	26.6	22.5	**25.0**	**28.0**	25.6	30.6
Doubao-1.5-Pro-32k (20250115)	-	**26.6**	**30.0**	22.5	**25.0**	25.6	**30.0**	24.1
Close Source MLLMs								
OpenAI-o3 (20250417)	40.1	29.5	24.2	31.0	34.6	27.0	25.0	41.9
GPT-4o-mini (20240718)	27.3	24.3	27.2	23.4	23.5	18.3	**31.1**	16.7
GPT-4o (20240806) [35]	30.0	26.3	28.6	24.7	27.2	26.8	20.0	25.9
Kimi-latest(202504) [69]	27.8	25.9	24.9	29.4	26.5	28.0	16.7	26.9
Doubao-1.6-Vision (250815)	**43.8**	**34.9**	**32.9**	**34.6**	**39.0**	**31.7**	30.0	**43.5**
Gemini-2.0-Pro (20250205) [68]	36.5	28.0	29.7	24.2	27.9	30.5	22.2	33.3
Claude-3.7-Sonnet (20250219)	33.5	24.8	22.7	27.3	27.9	28.0	22.2	22.2
Open Source MLLMs								
LLaVA-v1.5-7B [49]	25.3	24.6	26.1	24.2	23.5	17.1	31.1	22.2
LLaVA-OneVision-7B (SI) [41]	26.8	25.3	22.4	27.3	**33.1**	23.2	25.6	22.2
ShareGPT4V [12]	26.8	23.4	24.9	22.1	23.5	19.5	28.9	19.4
MiniCPM-o-2.6 [89]	28.8	25.3	25.6	23.0	27.3	21.9	24.5	29.9
GLM-4v-9B [27]	29.1	24.3	22.4	23.7	28.3	26.0	24.1	25.3
Ovis2-8B [56]	27.7	25.6	26.1	23.8	27.2	28.0	25.6	24.1
mPLUG-Owl3-7B-241101 [91]	25.6	18.9	21.5	15.2	16.2	20.7	18.9	20.4
Skywork-R1V3-38B [67]	31.2	**27.9**	26.5	**29.6**	24.6	21.2	26.6	**39.3**
Ernie-4-5-Turbo-VL [10]	31.0	27.1	27.8	25.0	24.5	**35.7**	30.0	24.4
Qwen2.5-VL-7B-Instruct [9]	30.1	26.0	27.6	20.9	25.2	23.2	**37.8**	25.0
Qwen2.5VL-72B-Instruct [9]	32.2	26.2	25.2	23.8	27.2	25.6	25.6	34.3
QvQ-72B-Preview [70]	29.8	23.0	24.2	17.0	24.4	21.0	24.4	30.6
InternVL2.5-38B [16]	33.3	25.5	24.4	26.4	27.2	23.2	25.6	26.9
InternVL2.5-78B [16]	30.7	27.3	26.6	26.0	26.5	26.8	31.1	30.6
InternVL3-38B [98]	33.2	27.1	**28.7**	27.6	26.1	21.4	23.9	28.5
InternVL3-78B [98]	**33.6**	27.7	27.7	26.1	31.6	26.3	21.3	32.3

based only on pure-text corpora; future works should explore CoT techniques tailored to multimodal data to better support visual reasoning tasks.

Effectiveness of Hint Prompts. Table 3 shows that hint prompts can boost model performance—*Claude-3.7-Sonnet*, *Gemini-2.0-Pro*, and *Doubao-1.5-Vision-Pro-32k* all improve by over 8 points, reaching accuracies above 35%. However, even with explicit guidance, models still fail to construct coherent, reliable reasoning chains. This suggests that simply augmenting training data with similar tasks is insufficient (which can help MLLMs come up with specific directions for solving the problem); future efforts must focus on enhancing the reliability and correctness of reasoning procedures of MLLMs to achieve more accurate reasoning inference. The complete results and analysis for all MLLMs can be found in Appendix C.

Impact of Model Scaling. In Table 1, we observe a positive correlation between parameter size and model performance. With in the same model series, *Qwen2.5-VL-72B-Instruct* achieves 26.2 % outperforming *Qwen2.5VL-7B-Instruct* (26.0%) by 0.2%. Furthermore, *InternVL2.5-78B* (27.3%) surpasses *InternVL2.5-38B* (25.5%) by a margin of 1.8%.

Open-Source vs Close-Source. Table 1 further compares open- and closed-source models. The top open-source MLLM, *InternVL3-78B*, attains 27.7%, trailing the closed-source leader (*OpenAI-o3*,

Table 2: **Influence of Chain-of-Thought on model performance.** Positive value changes are highlighted in red, negative changes in green, and statistically insignificant variations (delta < 1%) are denoted in gray. With CoT prompts, MLLMs only exhibit tiny improvements in visual reasoning.

Models	CoT	Overall	Quantity	Spatiality	Position	Attribute	Style	Other
GPT-4o (20240806)	✓	26.3	28.6	24.7	27.2	26.8	20.0	25.9
	✕	26.0$_{(-0.3)}$	26.9$_{(-1.7)}$	24.2$_{(-0.5)}$	26.5$_{(-0.7)}$	23.2$_{(-3.6)}$	24.0$_{(+4.0)}$	29.6$_{(+3.7)}$
Kimi-latest	✓	25.9	24.9	29.4	26.5	28.0	16.7	26.9
	✕	25.1$_{(-0.8)}$	22.9$_{(-2.0)}$	22.5$_{(-6.9)}$	25.0$_{(-1.5)}$	19.5$_{(-7.5)}$	35.6$_{(+18.9)}$	24.1$_{(-2.8)}$
GPT-4o-mini (20240718)	✓	24.3	27.2	23.4	23.5	18.3	31.1	16.7
	✕	23.1$_{(-1.2)}$	23.8$_{(-3.4)}$	22.9$_{(-0.5)}$	24.3$_{(+0.8)}$	17.1$_{(-1.2)}$	30.0$_{(-1.1)}$	18.5$_{(+1.8)}$
Qwen2.5-VL-Instruct-7B	✓	26.0	27.6	20.9	25.2	23.2	37.8	25.0
	✕	25.9$_{(-0.1)}$	25.5$_{(-2.1)}$	22.8$_{(+1.9)}$	26.4$_{(+1.2)}$	25.3$_{(+2.1)}$	20.6$_{(-17.2)}$	38.2$_{(+13.2)}$
InternVL2.5-38B	✓	24.9	24.1	26.4	27.2	23.2	25.6	22.2
	✕	25.0$_{(+0.1)}$	24.6$_{(+0.5)}$	25.5$_{(-0.9)}$	22.1$_{(-5.1)}$	22.0$_{(-1.2)}$	26.7$_{(+1.1)}$	29.6$_{(+7.4)}$

Table 3: **Influence of hint prompts on model performance.** MLLMs exhibit measurable performance enhancements with hint integration, yet retain significant gaps against human performance. In comparison, humans achieve task mastery on VisuLogic with hints. Value changes are color-coded with red indicating positive shifts and green denoting negative variations.

Models	Hint	Overall	Quantity	Spatiality	Position	Attribute	Style	Other
Human	✕	51.4	45.3	52.7	71.1	50.0	47.5	44.2
	✓	83.6$_{(+32.2)}$	85.1$_{(+39.8)}$	68.5$_{(+15.8)}$	100.0$_{(+28.9)}$	95.7$_{(+45.7)}$	78.6$_{(+31.1)}$	90.5$_{(+46.3)}$
GPT-4o (20240806)	✕	26.3	28.6	24.7	27.2	26.8	20.0	25.9
	✓	30.0$_{(+3.7)}$	25.4$_{(-3.2)}$	31.5$_{(+6.8)}$	29.2$_{(+2.0)}$	28.6$_{(+1.8)}$	30.8$_{(+10.8)}$	42.9$_{(+17.0)}$
Claude-3.7-Sonnet (20250219)	✕	24.8	22.7	27.3	27.9	28.0	22.2	22.2
	✓	33.5$_{(+8.7)}$	37.3$_{(+14.6)}$	33.3$_{(+6.0)}$	37.5$_{(+9.6)}$	23.8$_{(-4.2)}$	15.4$_{(-6.8)}$	38.1$_{(+15.9)}$
Gemini-2.0-Pro (20250205)	✕	28.0	29.7	24.2	27.9	30.5	22.2	33.3
	✓	36.5$_{(+8.5)}$	44.8$_{(+15.1)}$	33.3$_{(+9.1)}$	25.0$_{(-2.9)}$	38.1$_{(+7.6)}$	15.4$_{(-6.8)}$	42.9$_{(+9.6)}$
Doubao-1.5-Vision-Pro-32k (20250115)	✕	28.1	28.1	23.8	29.1	25.1	32.1	35.0
	✓	37.0$_{(+8.9)}$	46.3$_{(+18.2)}$	25.9$_{(+2.1)}$	54.2$_{(+25.1)}$	33.3$_{(+8.2)}$	23.1$_{(-9.0)}$	28.6$_{(-6.4)}$

29.5%) by only 1.8% points and outperforming other proprietary competitors such as *GPT-4o* and *Claude-3.7-Sonnet*. Overall, both open- and closed-source models exhibit uniformly low performance, highlighting a widespread neglect of visual reasoning objectives in current multimodal model training and data collection.

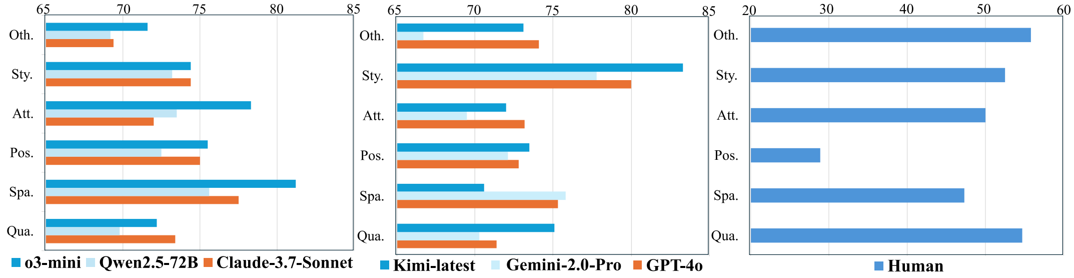
4.3 FINE-GRAINED COMPARISON

We systematically analyze model capabilities by examining error distributions across reasoning categories for different models. Figure 8 presents the error rates of LLMs, MLLMs, and human participants over six distinct reasoning categories.

Figure 8a reveals that LLMs struggle most with *Spatial Reasoning* questions, indicating that text-only descriptions are insufficient to infer three-dimensional structures or spatial transformations. In contrast, their performance on *Quantitative Reasoning* tasks is comparatively stronger, suggesting that quantitative relationships are more readily conveyed through language.

As shown in Figure 8b, *Stylistic Reasoning* presents the greatest difficulty for MLLMs, with error rates exceeding 75%—worse than random guessing (25% accuracy). This result underscores a fundamental limitation of current MLLM architectures in capturing subtle visual cues such as overlays, contours, and shape variations.

Figure 8c reveals that human error patterns form a distinct cluster, separate from LLMs and MLLMs. Human participants maintain error rates below 30% on *Positional Reasoning* tasks, reflecting robust position-based visual inference. In contrast, LLMs and MLLMs face significant challenges with positional reasoning, underscoring a fundamental divergence between human and model visual–cognitive processes and revealing limitations in how these models interpret positional information.



(a) LLMs' error distribution. (b) MLLMs' error distribution. (c) Humans' error distribution.

Figure 8: **Error distribution analysis.** The figure demonstrates distinct error type allocations across Humans, LLMs and MLLMs, revealing differences among their cognition patterns.

4.4 QUALITATIVE ANALYSIS

LLM Failures. As shown in Figure 7(h), LLMs that rely on externally generated image captions often omit critical visual details required for multi-step logical deduction—such as the counts, shapes, and progression patterns of the black and white dots in Figure 7(a). As a result, their reasoning deviates from the correct solution, often producing hallucinations or irrelevant responses.

MLLM Failures. Figure 7 also presents cases in which MLLMs correctly describe static visual content yet fail to infer the evolving relationships among shapes, instead resorting to superficial cues like object counts. While these models can recognize individual shapes and tally items, they struggle to reason over inter-element relations, which limits their ability to solve visual-logic problems. Besides, we provide failure mode analysis of each categories in VisuLogic in Appendix C.5.

5 CONCLUSION

In this paper, we present VisuLogic, a novel benchmark designed to evaluate the visual reasoning capabilities of Multi-modal Large Language Models (MLLMs). The benchmark consists of 1,000 vision-centric reasoning tasks distributed across six distinct categories. We conduct comprehensive evaluation of several advanced LLMs and MLLMs on this benchmark and provide an in-depth analysis of their performance. Our findings reveal that even the most advanced models fall short of human performance, highlighting substantial opportunities for advancement in visual logical reasoning. To promote further research and innovation, we'll open-source the evaluation code and datasets associated with this work. We hope this work serves as an important research in visual reasoning and contributes to the broader progress of MLLMs.

Reproducibility Statement. We provide detailed descriptions of dataset construction, preprocessing steps and quality controlling methods in Section 3 and Appendix B. Experimental setups, model configurations, and hyperparameters are reported in Section 4 and Appendix C. An anonymous link to download our dataset is included in the supplementary materials.

REFERENCES

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] W. U. Ahmad, S. Narenthiran, S. Majumdar, A. Ficek, S. Jain, J. Huang, V. Noroozi, and B. Ginsburg. Opencodereasoning: Advancing data distillation for competitive coding. *arXiv preprint arXiv:2504.01943*, 2025.

[3] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms, 2024.

[4] S. N. Akter, S. Lee, Y. Chang, Y. Bisk, and E. Nyberg. Visreas: Complex visual reasoning with unanswerable questions, 2024.

[5] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[6] R. An, S. Yang, M. Lu, R. Zhang, K. Zeng, Y. Luo, J. Cao, H. Liang, Y. Chen, Q. She, et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024.

[7] H. Bai, Y. Zhou, L. E. Li, S. Levine, and A. Kumar. Digi-q: Learning q-value functions for training device-control agents. *arXiv preprint arXiv:2502.15760*, 2025.

[8] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[9] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[10] Baidu-ERNIE-Team. Ernie 4.5 technical report, 2025.

[11] G. Chen, M. Wang, Y. Yang, K. Yu, L. Yuan, and Y. Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36:29667–29679, 2023.

[12] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[13] L. Chen, L. Li, H. Zhao, Y. Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. `https://github.com/Deep-Agent/R1-V`, 2025. Accessed: 2025-02-02.

[14] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023.

[15] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.

[16] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[17] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[18] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[19] S. Cheng, Z. Guo, J. Wu, K. Fang, P. Li, H. Liu, and Y. Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024.

[20] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[21] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

[22] H. Dong, J. Li, B. Wu, J. Wang, Y. Zhang, and H. Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[24] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.

[25] J. Feng, R. Xu, J. Hao, H. Sharma, Y. Shen, D. Zhao, and W. Chen. Language models can be logical solvers. *arXiv preprint arXiv:2311.06158*, 2023.

[26] Z. Gao, Z. Chen, E. Cui, Y. Ren, W. Wang, J. Zhu, H. Tian, S. Ye, J. He, X. Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024.

[27] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, H. Wang, J. Sun, J. Zhang, J. Cheng, J. Gui, J. Tang, J. Zhang, J. Li, L. Zhao, L. Wu, L. Zhong, M. Liu, M. Huang, P. Zhang, Q. Zheng, R. Lu, S. Duan, S. Zhang, S. Cao, S. Yang, W. L. Tam, W. Zhao, X. Liu, X. Xia, X. Zhang, X. Gu, X. Lv, X. Liu, X. Liu, X. Yang, X. Song, X. Zhang, Y. An, Y. Xu, Y. Niu, Y. Yang, Y. Li, Y. Bai, Y. Dong, Z. Qi, Z. Wang, Z. Yang, Z. Du, Z. Hou, and Z. Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

[28] B. Goertzel and C. Pennachin. *Artificial general intelligence*, volume 2. Springer, 2007.

[29] Z. Guo, R. Zhang, X. Zhu, Y. Tang, X. Ma, J. Han, K. Chen, P. Gao, X. Li, H. Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

[30] H. Gupta, S. Verma, U. Anantheswaran, K. Scaria, M. Parmar, S. Mishra, and C. Baral. Polymath: A challenging multi-modal mathematical reasoning benchmark, 2024.

[31] X. Han, Q. You, Y. Liu, W. Chen, H. Zheng, K. Mrini, X. Lin, Y. Wang, B. Zhai, J. Yuan, et al. Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models. *arXiv preprint arXiv:2311.11567*, 2023.

[32] Y. Hao, J. Gu, H. W. Wang, L. Li, Z. Yang, L. Wang, and Y. Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

[33] C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

[34] D. Huang, Q. Bu, Y. Qing, and H. Cui. Codecot: Tackling code syntax errors in cot reasoning for code generation. *arXiv preprint arXiv:2308.08784*, 2023.

[35] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[36] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[37] X. Jiang, Y. Dong, L. Wang, Z. Fang, Q. Shang, G. Li, Z. Jin, and W. Jiao. Self-planning code generation with large language models. *ACM Transactions on Software Engineering and Methodology*, 33(7):1–30, 2024.

[38] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset, 2017.

[39] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8888–8897, 2019.

[40] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. Lisa: Reasoning segmentation via large language model, 2024.

[41] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li. Llava-onevision: Easy visual task transfer, 2024.

[42] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[43] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[44] J. Li, G. Li, Y. Li, and Z. Jin. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23, 2025.

[45] J. Li, W. Lu, H. Fei, M. Luo, M. Dai, M. Xia, Y. Jin, Z. Gan, D. Qi, C. Fu, Y. Tai, W. Yang, Y. Wang, and C. Wang. A survey on benchmarks of multimodal large language models, 2024.

[46] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[47] W. Lin, X. Wei, R. An, P. Gao, B. Zou, Y. Luo, S. Huang, S. Zhang, and H. Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024.

[48] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[49] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023.

[50] H. Liu, Z. Teng, L. Cui, C. Zhang, Q. Zhou, and Y. Zhang. Logicot: Logical chain-of-thought instruction-tuning. *arXiv preprint arXiv:2305.12147*, 2023.

[51] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

[52] Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[53] Z. Liu, Z. Sun, Y. Zang, X. Dong, Y. Cao, H. Duan, D. Lin, and J. Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[54] Z. Liu, Y. Zhang, F. Liu, C. Zhang, Y. Sun, and J. Wang. Othink-mr1: Stimulating multimodal generalized reasoning capabilities through dynamic reinforcement learning. *arXiv preprint arXiv:2503.16081*, 2025.

[55] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[56] S. Lu, Y. Li, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and H.-J. Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024.

[57] K. Mangalam, R. Akshulakov, and J. Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[58] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[59] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[60] M. Mathew, D. Karatzas, and C. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[61] F. Meng, L. Du, Z. Liu, Z. Zhou, Q. Lu, D. Fu, B. Shi, W. Wang, J. He, K. Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[62] T. Nguyen, S. Y. Gadre, G. Ilharco, S. Oh, and L. Schmidt. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36:22047–22069, 2023.

[63] R. Niu, J. Li, S. Wang, Y. Fu, X. Hu, X. Leng, H. Kong, Y. Chang, and Q. Wang. Screenagent: A vision language model-driven computer control agent. *arXiv preprint arXiv:2402.07945*, 2024.

[64] Y. Peng, G. Zhang, M. Zhang, Z. You, J. Liu, Q. Zhu, K. Yang, X. Xu, X. Geng, and X. Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[65] R. Qiao, Q. Tan, G. Dong, M. Wu, C. Sun, X. Song, Z. GongQue, S. Lei, Z. Wei, M. Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

[66] H. Shen, Z. Zhang, K. Zhao, Q. Zhang, R. Xu, and T. Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. `https://github.com/om-ai-lab/VLM-R1`, 2025. Accessed: 2025-02-15.

[67] W. Shen, J. Pei, Y. Peng, X. Song, Y. Liu, J. Peng, H. Sun, Y. Hao, P. Wang, J. Zhang, and Y. Zhou. Skywork-r1v3 technical report, 2025.

[68] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[69] K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[70] Q. Team. Qvq: To see the world with wisdom, December 2024.

[71] Q. Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025.

[72] Y. Wan, W. Wang, Y. Yang, Y. Yuan, J.-t. Huang, P. He, W. Jiao, and M. R. Lyu. Logicasker: Evaluating and improving the logical reasoning ability of large language models. *arXiv preprint arXiv:2401.00757*, 2024.

[73] K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, and H. Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[74] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[75] S. Wang, D. Kim, A. Taalimi, C. Sun, and W. Kuo. Learning visual grounding from generative vision and language model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8057–8067. IEEE, 2025.

[76] W. Wang, Z. Gao, L. Chen, Z. Chen, J. Zhu, X. Zhao, Y. Liu, Y. Cao, S. Ye, X. Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.

[77] X. Wang and P. Peng. Open-r1-video. `https://github.com/Wang-Xiaodong1899/Open-R1-Video`, 2025.

[78] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning, 2024.

[79] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[80] Y. Xiao, E. Sun, T. Liu, and W. Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[81] Z. Xie and C. Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024. *URL https://arxiv. org/abs/2408.16725*, 2024.

[82] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[83] F. Xu, Z. Wu, Q. Sun, S. Ren, F. Yuan, S. Yuan, Q. Lin, Y. Qiao, and J. Liu. Symbol-llm: Towards foundational symbol-centric interface for large language models. *arXiv preprint arXiv:2311.09278*, 2023.

[84] R. Xu, Z. Huang, T. Wang, Y. Chen, J. Pang, and D. Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024.

[85] Y. Xu, X. Liu, X. Liu, Z. Hou, Y. Li, X. Zhang, Z. Wang, A. Zeng, Z. Du, W. Zhao, et al. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. *arXiv preprint arXiv:2404.02893*, 2024.

[86] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[87] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[88] Y. Yang, X. He, H. Pan, X. Jiang, Y. Deng, X. Yang, H. Lu, D. Yin, F. Rao, M. Zhu, B. Zhang, and W. Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[89] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[90] J. Ye, G. Li, S. Gao, C. Huang, Y. Wu, S. Li, X. Fan, S. Dou, Q. Zhang, T. Gui, et al. Tooleyes: fine-grained evaluation for tool learning capabilities of large language models in real-world scenarios. *arXiv preprint arXiv:2401.00741*, 2024.

[91] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024.

[92] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[93] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

[94] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[95] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, Y. Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[96] Q. Zhao, S. Wang, C. Zhang, C. Fu, M. Q. Do, N. Agarwal, K. Lee, and C. Sun. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*, 2023.

[97] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[98] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, Y. Duan, H. Tian, W. Su, J. Shao, Z. Gao, E. Cui, Y. Cao, Y. Liu, X. Wei, H. Zhang, H. Wang, W. Xu, H. Li, J. Wang, D. Chen, S. Li, Y. He, T. Jiang, J. Luo, Y. Wang, C. He, B. Shi, X. Zhang, W. Shao, J. He, Y. Xiong, W. Qu, P. Sun, P. Jiao, H. Lv, L. Wu, K. Zhang, H. Deng, J. Ge, K. Chen, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

APPENDIX

The Use of LLMs. We only used LLMs to polish the paper and did not involve them in the core contributions of this work.

A OVERVIEW OF THE APPENDIX

In the appendix, we provide additional details and supplementary information to further elaborate on sections mentioned above. In Section B, we analyze the statistical features of the dataset, meanwhile providing examples of questions ranging from different categories. Section C contains experiments details, including the evaluation of LLMs, the evaluation of hint prompts and RL experiments. Some examples of model outputs are also illustrated.

Due to file size limitations, we place all supplementary materials related to the paper in an anonymous link https://anonymous.4open.science/r/4644a0d29c4ded212c057467c54df6d5. The anonymous repository contains the benchmark data used in the paper, along with an additional training dataset.

B BENCHMARK ANALYSIS

B.1 STATISTICAL ANALYSIS

As shown in Figure 10, the text length of questions in VisuLogic is mostly concentrated around 40 tokens (calculated by Llama-3.1's and InternVL2.5's tokenizer). We also analyze the distribution of image sizes, as shown in Figure 9. The image widths range from 200 to 700 pixels, with an average of 592.3 pixels, while the heights range from 90 to 825 pixels, with an average of 327.9 pixels.

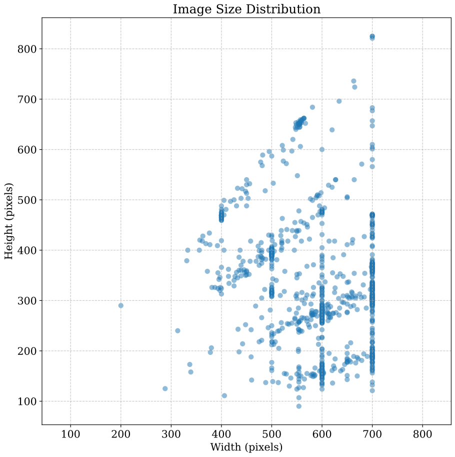


Figure 9: **Image size distribution.** The size of images is limited to within the same order of magnitude.

B.2 MORE EXAMPLES OF VISULOGIC

To provide a thoroughly presentation of our benchmark, we include more examples of questions from different categories in the Figure 11 and Figure 12.

B.3 TRAINING DATASET

To facilitate further investigation of visual reasoning, we provide an auxiliary training set of 4,296 question–answer pairs drawn from the same domains and subjected to identical validation procedures

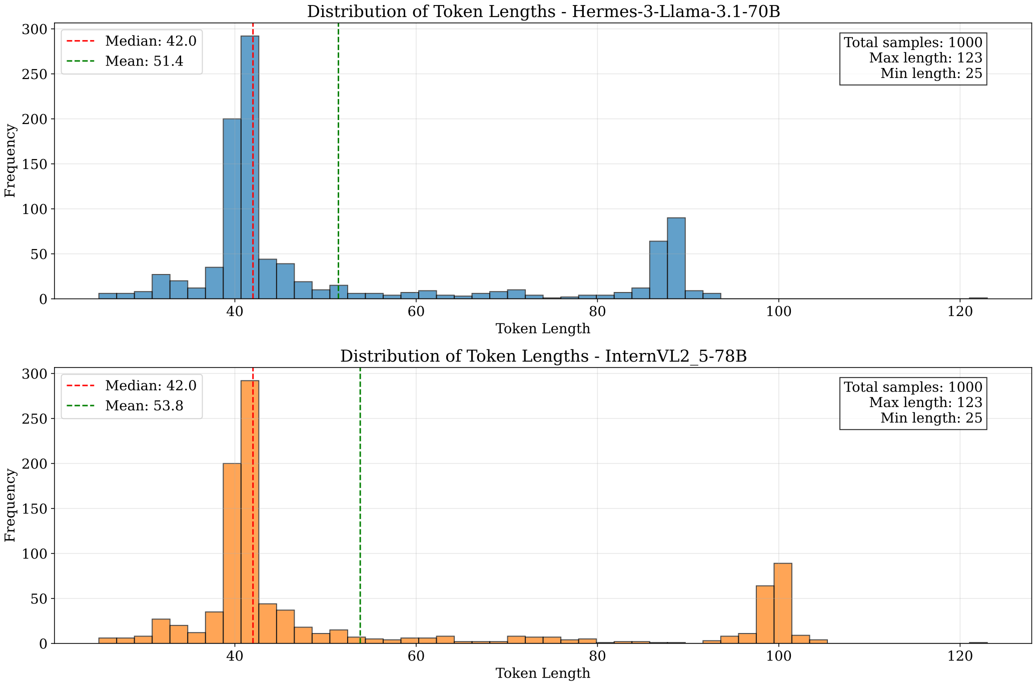


Figure 10: Distribution of text token length in VisuLogic.

to prevent overlap with the benchmark. The training split mirrors the primary taxonomy, with category proportions of Quantitative Reasoning (30.7%), Spatial Reasoning (25.5%), Positional Reasoning (13.0%), Attribute Reasoning (8.8%), Stylistic Reasoning (9.9%), and Other (12.1%).

B.4 ETHICAL CONSIDERATIONS

Data Source. Our data is sourced from `https://www.fenbi.com/spa/tiku`. The website's robots.txt `https://www.fenbi.com/robots.txt` file permits crawling.

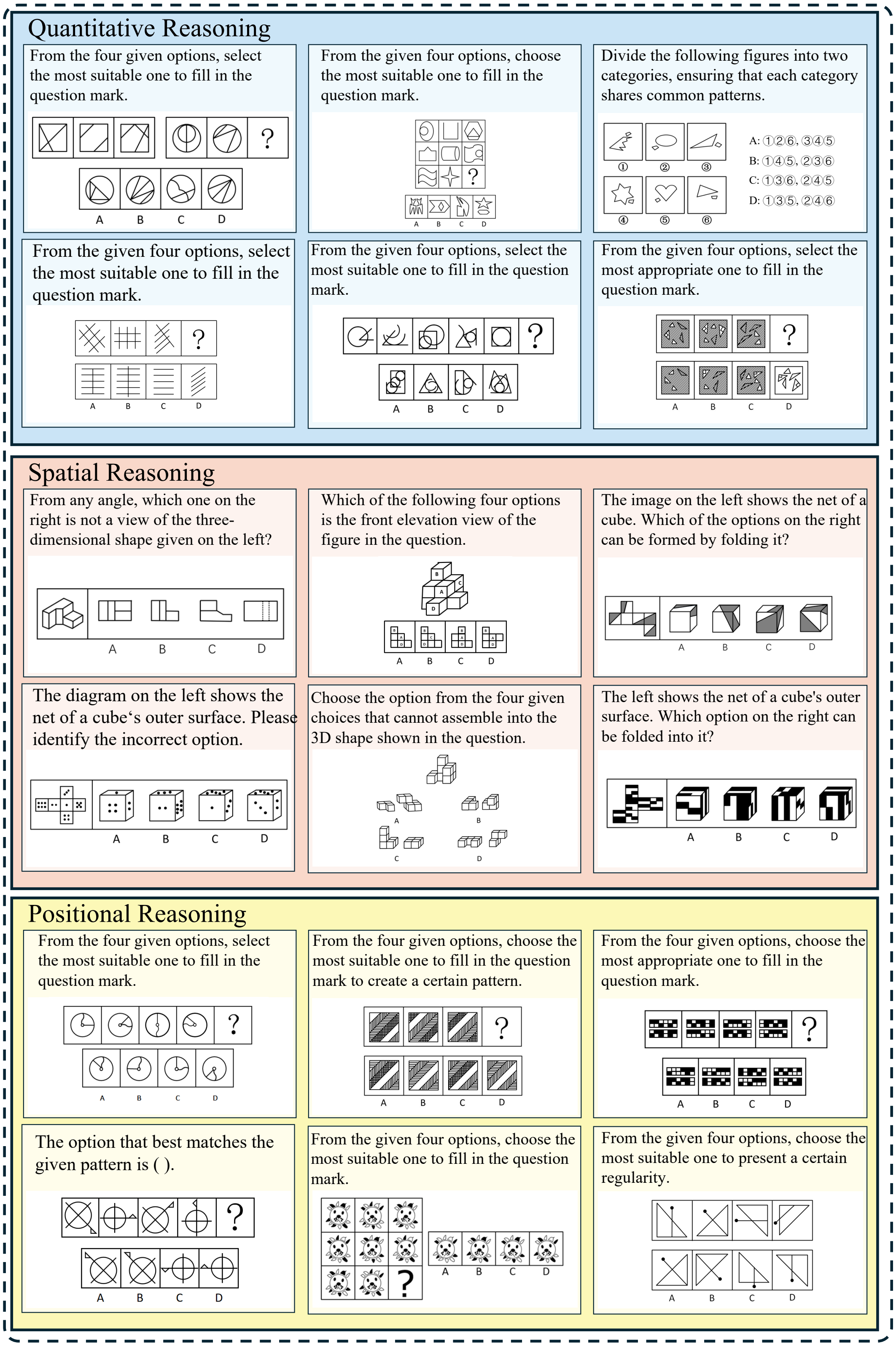
Ethics of Human Data Annotation.

1. **Transparency.** All potential annotators received a detailed briefing outlining the specific nature, expected outcomes, and necessary skills for the annotation tasks prior to commencing work. We provided all annotators with information regarding the usage of the data, as well as whether or not it will be publicly released.

2. **Fair Compensation.** We implemented an equitable compensation scheme—clearly defining all payment terms upfront, ensuring annotators were paid well above local minimum-wage standards, and issuing verified payments promptly based on tracked task-completion times.

3. **Privacy Protection.** We strictly anonymized annotators' PII at recruitment, presented only fully screened and de-identified source data during annotation, and restricted all collected annotation data and metadata to the core research team to ensure security and confidentiality.

This study is designed to be non-invasive, presenting no physical risk to the subjects. The tasks performed are benign [e.g., viewing stimulus material, answering multiple-choice questions, or interacting with a standard user interface]. We have reviewed the content to ensure it is not offensive, will not cause emotional distress, and will not induce stress or anxiety. Therefore, we can confirm that this study poses no adverse mental or physical effects on the subjects. Hence, we conclude that this study is exempt from IRB review.



Quantitative Reasoning

From the four given options, select the most suitable one to fill in the question mark.

From the given four options, choose the most suitable one to fill in the question mark.

Divide the following figures into two categories, ensuring that each category shares common patterns.

From the given four options, select the most suitable one to fill in the question mark.

From the given four options, select the most suitable one to fill in the question mark.

From the given four options, select the most appropriate one to fill in the question mark.

Spatial Reasoning

From any angle, which one on the right is not a view of the three-dimensional shape given on the left?

Which of the following four options is the front elevation view of the figure in the question.

The image on the left shows the net of a cube. Which of the options on the right can be formed by folding it?

The diagram on the left shows the net of a cube‘s outer surface. Please identify the incorrect option.

Choose the option from the four given choices that cannot assemble into the 3D shape shown in the question.

The left shows the net of a cube's outer surface. Which option on the right can be folded into it?

Positional Reasoning

From the four given options, select the most suitable one to fill in the question mark.

From the four given options, choose the most suitable one to fill in the question mark to create a certain pattern.

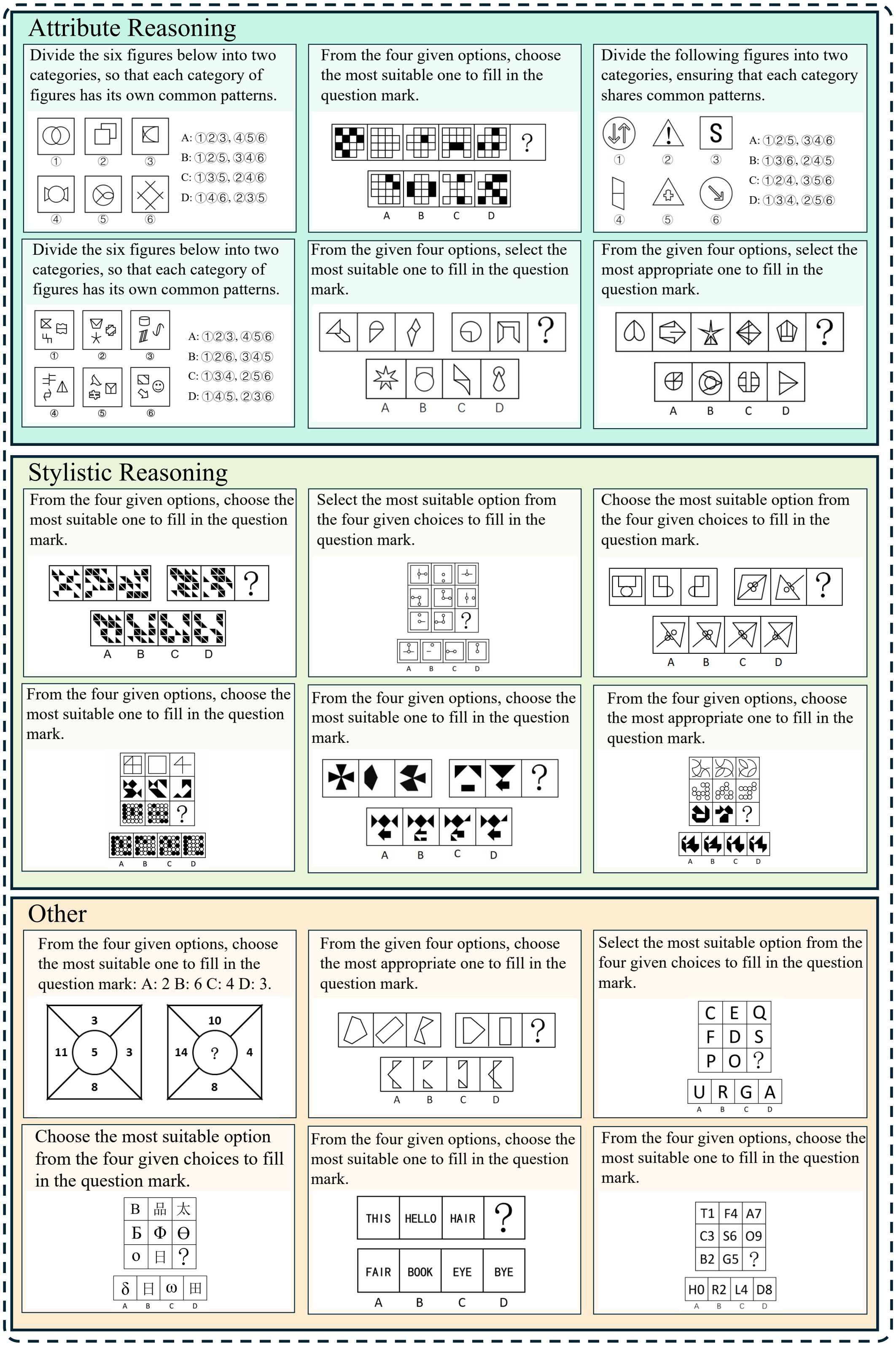
From the four given options, choose the most appropriate one to fill in the question mark.

The option that best matches the given pattern is ( ).

From the given four options, choose the most suitable one to fill in the question mark.

From the given four options, choose the most suitable one to present a certain regularity.

Figure 11: More examples in VisuLogic of Quantitative Reasoning, Spatial Reasoning, Positional Reasoning.



Attribute Reasoning

Divide the six figures below into two categories, so that each category of figures has its own common patterns.

From the four given options, choose the most suitable one to fill in the question mark.

Divide the following figures into two categories, ensuring that each category shares common patterns.

Divide the six figures below into two categories, so that each category of figures has its own common patterns.

From the given four options, select the most suitable one to fill in the question mark.

From the given four options, select the most appropriate one to fill in the question mark.

Stylistic Reasoning

From the four given options, choose the most suitable one to fill in the question mark.

Select the most suitable option from the four given choices to fill in the question mark.

Choose the most suitable option from the four given choices to fill in the question mark.

From the four given options, choose the most suitable one to fill in the question mark.

From the four given options, choose the most suitable one to fill in the question mark.

From the four given options, choose the most appropriate one to fill in the question mark.

Other

From the four given options, choose the most suitable one to fill in the question mark: A: 2 B: 6 C: 4 D: 3.

From the given four options, choose the most appropriate one to fill in the question mark.

Select the most suitable option from the four given choices to fill in the question mark.

Choose the most suitable option from the four given choices to fill in the question mark.

From the four given options, choose the most suitable one to fill in the question mark.

From the four given options, choose the most suitable one to fill in the question mark.

Figure 12: More examples in VisuLogic of Attribute Reasoning, Stylistic Reasoning, and Other.

C EVALUATION & EXPERIMENT

C.1 COST OF HUMAN EVALUATION

Because VisuLogic relies on a heavily automated pipeline—syntactic checks, image–caption consistency filters, and rule-based unit tests—only a single, lightweight human pass is needed at the end. Concretely, the entire benchmark (4296 train + 1000 test items) was reviewed by five annotators working two hours each (around 10 person-hours total). At a typical crowd-platform rate of USD 120, or \$0.02 per item.

C.2 EVALUATION OF LLMS

Caption Generation for LLMs Evaluation. In our experiment, we employ large language models (LLMs) for comparative analysis. Specifically, when setting up the LLM-based experiment, we initially utilize GPT-4o to generate captions for images with the following prompt: *Please describe the fine-grained content of the image or figure based on this question, including scenes, objects, relationships, and any text present. Please note that you do not need to answer this question directly, just describe the information of this picture*. Additional examples of generated image captions are presented in Figure 14 and Figure 15.

More Examples of Captions. We provide additional image captions for six categories, as illustrated in Figures 14 and 15. Even SOTA MLLM (GPT-4o) encounters difficulties in accurately describing the details of images from VisuLogic.

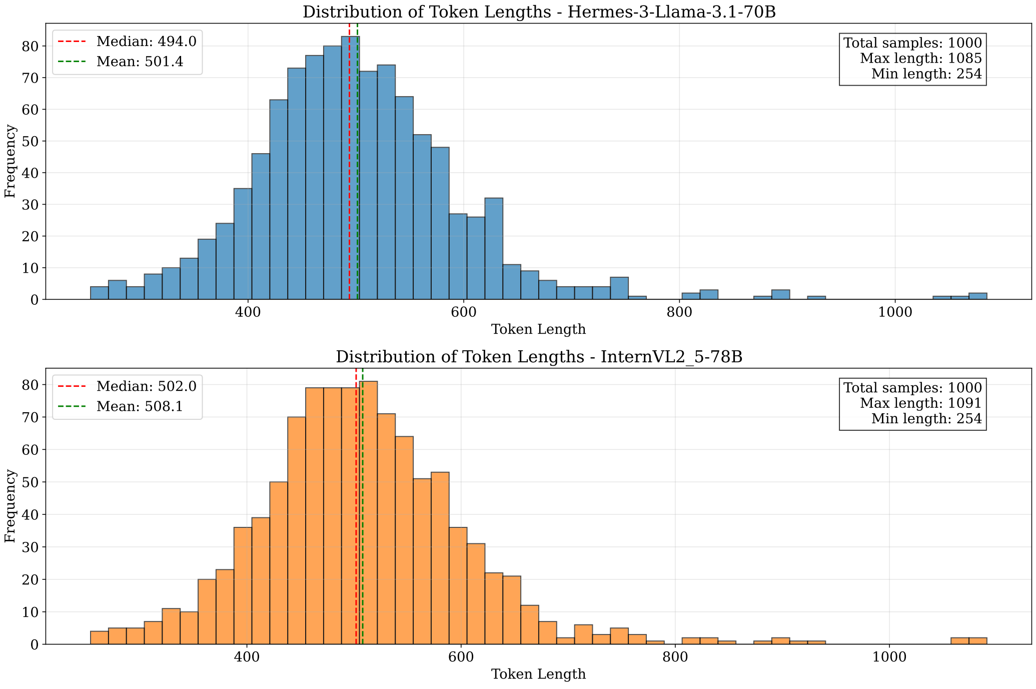


Figure 13: Distribution of tokens length in LLM evaluation settings, including image description.

Evaluation of Caption Quality Generated by MLLMs. We conducted a human evaluation of four MLLMs (GPT-4o, Claude Sonnet 3.5, Gemini 2.5 Flash, and Qwen2.5-VL-72B) across six dimensions of caption quality. Using 100 image-pair questions, we collected a total of 400 captions and invited 50 human evaluators to assess them. Each caption was independently evaluated by four annotators, with tasks distributed evenly to ensure balanced coverage across models. On average, each annotator evaluated 32 captions, corresponding to eight captions from each of the four models. Caption assignments were randomized to minimize systematic bias. The final evaluation results were obtained by averaging the scores for each model on each dimension.

Table 4: The average standard deviation of scores given by different evaluators for GPT-4o.

	Integrity	**Granularity**	**Hierarchy**	**Task Alignment**	**Clarity**	**Reasoning**
mean	0.266	0.307	0.324	0.328	0.326	0.582

Table 5: Detailed scores of captions from different models across various evaluation dimensions.

Model	**Integrity**	**Granularity**	**Hierarchy**	**Task Alignment**	**Clarity**	**Reasoning**
GPT-4o	1.78 (0.41)	1.75 (0.43)	1.73 (0.44)	1.74 (0.44)	1.71 (0.45)	0.99 (0.72)
Claude-Sonnet-3.5	1.00 (0.74)	1.01 (0.69)	1.78 (0.41)	0.95 (0.69)	1.74 (0.44)	0.26 (0.44)
Gemini-2.5-Flash	1.04 (0.70)	0.24 (0.43)	0.96 (0.71)	0.24 (0.43)	0.98 (0.67)	0.25 (0.43)
Qwen2.5-VL-72B	1.74 (0.44)	0.91 (0.67)	1.76 (0.43)	1.04 (0.70)	1.00 (0.72)	1.70 (0.46)

Table 5 presents the results. We adopt a three-level scoring scheme (high = 2, mid = 1, low = 0) and report both the mean score and the variance to reflect annotator consistency. The table lists unrounded mean scores, with standard deviations shown in parentheses.

The six evaluation dimensions are defined as follows: 1) *Integrity*: whether the description comprehensively covers the main image content without omitting key elements. 2) *Granularity*: the level of detail in the description, capturing how specific and fine-grained the information is. 3) *Hierarchy*: the organization and structure of the description, including clarity of main points and subordinate details. 4) *Task Alignment*: how well the description aligns with the given task requirements or instructions. 5) *Clarity*: the readability and understandability of the language, emphasizing conciseness and avoidance of ambiguity. 6) *Reasoning*: whether the description demonstrates logical reasoning and causal interpretation of the image content.

Besides, we also conduct experiments to measure inter-annotator agreement among evaluators. For GPT-4o's "inter-annotator agreement among evaluators", we also use the standard deviation (std) as the metric. Specifically, we calculate the std of evaluator scores on each caption for every dimension, and then compute the average of these std values to quantify agreement. Notably, the standard deviation (std) here differs from the previously mentioned std of all scores. In this case, the std is calculated across different evaluators' ratings for the same caption and the same evaluation dimension.

As Table 4 shows, the average standard deviation of scores given by different evaluators remains low, illustrating high inter-annotator agreement for our evaluation.

C.3 MORE SOLUTIONS FROM MODELS

We provide more solutions generated from different LLMs/MLLMs on our benchmark, as shown in Figure 16, Figure 17 and Figure 18. For the majority of questions, almost all models fail to provide accurate solutions. Sometimes even when the final answer is correct, methodological wrong may persist.

C.4 VERIFICATION OF BENCHMARK SUFFICIENCY

In our evaluation of GPT-4o on the VisuLogic benchmark, we tested three sets of problems at each step. As we increased the number of benchmark questions from 10 to 1280, the standard deviation of the accuracy scores (corresponding to a 68.27% confidence interval) changed as follows:

The result shows that with around 1000 benchmark questions, the standard deviation has reached approximately 0.006, which is sufficient to reliably evaluate the model's capability in this aspect.

C.5 MORE FAILURE-MODE ANALYSIS

In Section 4, we analyze some failure issues. In this part we provide more detailed analysis for each categories. For commonly used models such as GPT-4o and Gemini 2.0 Pro, we have summarized the issues in Table 7.

Table 6: Stability of GPT-4o accuracy on the VisuLogic benchmark as sample size grows.

Index	Number of Questions	Std. Dev.
0	10	0.135
1	20	0.094
2	40	0.065
3	80	0.037
4	160	0.036
5	320	0.029
6	640	0.010
7	1000	0.0064
8	1280	0.0052

Table 7: Failure-mode analysis of MLLMs across question categories on the VisuLogic benchmark.

Categories	Analysis
Position Reasoning	(1) The pattern descriptions are often vague, using phrases like "in some way" or "following a certain rule," which hinders further reasoning. (2) Even when identifying correct patterns, the explanations lack specificity. For instance, saying "the circle moves up by one grid" doesn't capture the full logic when the grid itself may carry structural or relational significance.
Spatial Reasoning	(1) The contents on each face are only roughly described, and the model often treats distinct faces as similar. It struggles with complex shapes on individual faces. (2) When shapes are slanted or irregular, the model fails to match the unfolded and folded views of a 3D object. (3) Spatial reasoning techniques are mostly limited to face adjacency, lacking more advanced spatial transformation abilities.
Attribute Reasoning	(1) Uses imprecise descriptors like "more edges," "more protruding parts," or "more complex shapes," which are not clearly defined or measurable, making verification and comparison difficult. (2) Shows little evidence of tracking dynamic changes across a sequence; instead, it tends to only compare two figures at a time. (3) Fails to recognize changes in symmetry axes, showing limited sensitivity to symmetry transformations.
Quantitative Reasoning	(1) Uses abstract phrases like "more lines" or "becomes nested/symmetrical," without quantifying these changes (e.g., how many lines were added or what structure was formed). (2) Some conclusions, like 'complex → simple → complex,' are based on subjective impressions rather than measurable criteria. (3) Unable to analyze figures that do not consist of basic elements; often resorts to describing only vague or similar shapes.
Stylistic Reasoning	(1) In tasks involving rotation, it might state that "rotation is involved," but fails to specify the rotation angle, whether there is central symmetry, or which figures are related through rotation. (2) When dealing with complex visual styles, the reasoning process becomes disjointed and lacks clear correspondence between described features and actual image elements. (3) Multiple distinct patterns in a task hinder the model's ability to generalize consistent stylistic rules.
Other Reasoning	(1) For letter or character-based tasks, the model is often misled by the semantic meaning of the symbols, which interferes with pattern recognition and hinders reflective correction. (2) When facing mixed-type symbol patterns, the model has difficulty synthesizing rules across different symbol categories.

C.6 HINT PROMPTS EVALUATION DETAILS

We first generate hint prompts with GPT-4o, combining reference solutions with question data as inputs (see Figure 6). All outputs undergo manual validation to prevent solution leakage. More examples are shown in Figure 19. After that, we input the hint prompts along with the same CoT prompt in CoT experiments ("*Solve the complex visual logical reasoning problem through step-by-step reasoning. Think about the reasoning process first and answer the question following this format: Answer: \boxed{$LETTER}*.") to MLLMs.

Table 8: **Results of hint prompts for all models.** Value changes are color-coded with red indicating positive shifts and green denoting negative variations.

Models	Hint	Overall	Quantity	Spatiality	Position	Attribute	Style	Other
Human	✗	51.4	45.3	52.7	71.1	50.0	47.5	44.2
	✓	83.6 (+32.2)	85.1 (+39.8)	68.5 (+15.8)	100.0 (+28.9)	95.7 (+45.7)	78.6 (+31.1)	90.5 (+46.3)
OpenAI-o3 (20250417)	✗	29.5	24.2	31.0	34.6	27.0	25.0	41.9
	✓	40.1 (+10.6)	43.4 (+19.2)	33.0 (+2.0)	33.3 (−1.3)	54.6 (+27.6)	36.0 (+11.0)	37.5 (−4.4)
GPT-4o-mini (20240718)	✗	24.3	27.2	23.4	23.5	18.3	31.1	16.7
	✓	27.3 (+3.0)	24.7 (−2.5)	34.6 (+11.2)	27.9 (+4.4)	22.0 (+3.7)	24.4 (−6.7)	25.9 (+9.2)
GPT-4o (20240806)	✗	26.3	28.6	24.7	27.2	26.8	20.0	25.9
	✓	30.0 (+3.7)	25.4 (−3.2)	31.5 (+6.8)	29.2 (+2.0)	28.6 (+1.8)	30.8 (+10.8)	41.9 (+16.0)
Kimi-latest (202504)	✗	25.9	24.9	29.4	26.5	28.0	16.7	26.9
	✓	27.8 (+1.9)	27.5 (+2.6)	27.3 (−2.1)	19.9 (−6.6)	32.9 (+4.9)	26.7 (+10.0)	37.0 (+10.1)
Doubao-1.6-Vision (250815)	✗	34.9	32.9	34.6	39.0	31.7	30.0	43.5
	✓	43.8 (+8.9)	45.6 (+12.7)	39.8 (+5.2)	41.9 (+2.9)	50.0 (+18.3)	34.4 (+4.4)	51.9 (+8.4)
Gemini-2.0-Pro (20250205)	✗	28.0	29.7	24.2	27.9	30.5	22.2	33.3
	✓	36.5 (+8.5)	44.8 (+15.1)	33.3 (+9.1)	25.0 (−2.9)	38.1 (+7.6)	15.4 (−6.8)	42.9 (+9.6)
Claude-3.7-Sonnet (20250219)	✗	24.8	22.7	27.3	27.9	28.0	22.2	22.2
	✓	33.5 (+8.7)	37.3 (+14.6)	33.3 (+6.0)	37.5 (+9.6)	23.8 (−4.2)	15.4 (−6.8)	38.1 (+15.9)
LLaVA-v1.5-7B	✗	24.6	26.1	24.2	23.5	17.1	31.1	22.2
	✓	25.3 (+0.7)	26.6 (+0.5)	23.4 (−0.8)	24.3 (+0.8)	18.3 (+1.2)	37.8 (+6.7)	21.3 (−0.9)
LLaVA-OneVision-7B (SI)	✗	25.3	22.4	27.3	33.1	23.2	25.6	22.2
	✓	26.8 (+1.5)	27.5 (+5.1)	29.4 (+2.1)	25.0 (−8.1)	19.5 (−3.7)	28.9 (+3.3)	25.0 (+2.8)
ShareGPT4V	✗	23.4	24.9	22.1	23.5	19.5	28.9	19.4
	✓	26.7 (+3.3)	27.8 (+2.9)	29.4 (+7.3)	20.6 (−2.9)	23.2 (+3.7)	26.7 (−2.2)	27.8 (+8.4)
MiniCPM-o-2.6	✗	25.3	25.6	23.0	27.3	21.9	24.5	29.9
	✓	28.8 (+3.5)	28.1 (+2.5)	31.6 (+8.6)	26.5 (−0.8)	34.2 (+12.3)	24.4 (−0.1)	27.8 (−2.1)
GLM-4v-9B	✗	24.3	22.4	23.7	28.3	26.0	24.1	25.3
	✓	29.1 (+4.8)	24.1 (+1.7)	30.3 (+6.6)	38.2 (+9.9)	29.3 (+3.3)	25.6 (+1.5)	34.3 (+9.0)
Ovis2-8B	✗	25.6	26.1	23.8	27.2	28.0	25.6	24.1
	✓	27.7 (+2.1)	29.2 (+3.1)	29.9 (+6.1)	23.5 (−3.7)	29.3 (+1.3)	23.3 (−2.3)	25.9 (+1.8)
mPLUG-Owl3-7B-241101	✗	18.9	21.5	15.2	16.2	20.7	18.9	20.4
	✓	25.6 (+6.7)	27.8 (+6.3)	27.7 (+12.5)	20.6 (+4.4)	25.6 (+4.9)	15.6 (−3.3)	29.6 (+9.2)
Skywork-R1V3-38B	✗	27.9	26.5	29.6	24.6	21.2	26.6	39.3
	✓	31.2 (+3.3)	30.9 (+4.4)	30.7 (+1.1)	25.4 (+0.8)	36.7 (+15.5)	33.3 (+6.7)	35.3 (−4.0)
Ernie-4-5-Turbo-VL	✗	27.1	27.8	25.0	24.5	35.7	30.0	24.4
	✓	31.0 (+3.9)	28.6 (+0.8)	32.0 (+7.0)	27.2 (+2.7)	35.4 (−0.3)	32.2 (+2.2)	37.0 (+12.6)
Qwen2.5-VL-7B-Instruct	✗	26.0	27.6	20.9	25.2	23.2	37.8	25.0
	✓	30.1 (+4.1)	34.0 (+6.4)	29.4 (+8.5)	25.0 (−0.2)	20.7 (−2.5)	22.2 (−15.6)	38.9 (+13.9)
Qwen2.5-VL-72B-Instruct	✗	26.2	25.2	23.8	27.2	25.6	25.6	34.3
	✓	32.2 (+6.0)	32.3 (+7.1)	35.1 (+11.3)	25.0 (−2.2)	28.1 (+2.5)	34.4 (+8.8)	37.0 (+2.7)
QvQ-72B-Preview	✗	23.0	24.2	17.0	24.4	21.0	24.4	30.6
	✓	29.8 (+6.8)	30.1 (+5.9)	39.7 (+22.7)	19.6 (−4.8)	34.2 (+13.2)	16.7 (−7.7)	34.0 (+3.4)
InternVL2.5-38B	✗	25.5	24.4	26.4	27.2	23.2	25.6	26.9
	✓	33.3 (+7.8)	32.9 (+8.5)	36.8 (+10.4)	34.56 (+7.4)	31.7 (+8.5)	31.1 (+5.5)	26.9 (+0.0)
InternVL2.5-78B	✗	27.3	26.6	26.0	26.5	26.8	31.1	30.6
	✓	30.7 (+3.4)	30.3 (+3.7)	34.6 (+8.6)	27.2 (+0.7)	34.2 (+7.4)	26.7 (−4.4)	29.6 (−1.0)
InternVL3-38B	✗	27.1	28.7	27.6	26.1	21.4	23.9	28.5
	✓	33.2 (+6.1)	35.4 (+6.7)	39.0 (+11.4)	24.3 (−1.8)	25.6 (+4.2)	24.4 (+0.5)	38.0 (+9.5)
InternVL3-78B	✗	27.7	27.7	26.1	31.6	26.3	21.3	32.3
	✓	33.6 (+5.9)	35.7 (+8.0)	35.9 (+9.8)	27.2 (−4.4)	32.9 (+6.6)	27.8 (+6.5)	35.2 (+2.9)

Table 8 presents the complete results of the hint experiments.

Overall Trends. Most MLLMs improve with hints, but gains are modest (≈3–10 points) and substantially smaller than for humans. Several strong positives include *Doubao-1.6-Vision* (34.9→43.8; +8.9), *Claude-3.7-Sonnet* (24.8→33.5; +8.7), and the *InternVL* family (*e.g.*, InternVL2.5-38B: 25.5→33.3; +7.8). In contrast, *Kimi-latest* shows limited net change (25.9→27.8; +1.9). This indicates that models differ markedly in their reasoning abilities.

Dimension-wise Effects. Gains are uneven across skills. *Attribute* and *Other* often show the largest improvements (e.g., Doubao: +18.3 on *Attribute*; Claude: +15.9 on *Other*). *Quantity* also improves notably for several models (e.g., Gemini: +15.1; Claude: +14.6; InternVL lines: +6–9). However, we also observe regressions: *Position* degrades for some models (e.g., Kimi: −6.6; GPT-4o-mini: +4.4 but with other dips), and *Style* can suffer negative transfer (e.g., Qwen2.5-VL-7B: −15.6; Gemini: −6.8; Claude: −6.8). The results indicate that models vary in their specializations, with certain models showing deficiencies in specific domains.

Model-Specific Analysis. *GPT-4o* and *GPT-4o-mini* exhibit small overall gains (+3.7 and +3.0, respectively), with mixed per-dimension changes (e.g., improvements on *Spatiality* but dips on *Quantity* and *Style*). *OpenAI-o3* shows a moderate overall lift (+10.6) driven by *Quantity* and *Attribute*, yet small declines on *Position* and *Other*. Larger InternVL models (*e.g.*, InternVL3-78B) improve more consistently across *Quantity*, *Spatiality* and *Attribute*.

Concluding Remarks. (1) Although hints provide substantial help for humans, they yield only limited and inconsistent gains for current MLLMs, leaving a large human–model gap. (2) Improvements concentrate on *Attribute*, *Other*, and *Quantity*, while *Position* and *Style* can regress, suggesting incomplete exploitation of models' abilities.

C.7 RL EXPERIMENTS

We further include two reinforcement-learning baselines built on *Qwen2.5-VL-7B-Instruct* [9] and *InternVL2.5-38B* [18], respectively, trained via our rule-based RL procedure on our supplementary training dataset. Fully supervised fine-tuning (SFT) experiments on the same datasets serve as controls to isolate the effect of RL optimization. **Comparative SFT Experiments.** To verify the effectiveness of RL method, we arrange the comparative SFT experiments on the same dataset as RL experiments. The instruction consists of questions and Non-CoT prompts, and the responses are formatted direct answers.

RL Algorithm. We employ REINFORCE Leave-One-Out (RLOO) [3] in our reinforcement learning training phase. As a critic-model-free algorithm,rloo is at a low computational cost while maintaining more robustness to noise and KL constraints.

Reward Modeling. Inspired by Deepseek-R1 [20], we design our rule-based reward system that mainly consists of two types of rewards:

1. **Format rewards:** To clarify model's outputs, we design a format rule that forces model to put its thinking process between '<think>' and '</think>' tags and put its final answer between '<answer>' and '</answer>' tags. Regular expression is applied to judge whether outputs conform to the format rule.
2. **Accuracy rewards:** The accuracy reward is decided by the response's correctness. The model should generate the response in right format, then the answer is extracted and judged whether it is matched to the correct option.

Hyperparameter settings. Our two RL models are trained with the hyperparameter configuration detailed in Table 10. And the hyperparameters used in SFT training stage are listed in Table 9.

Table 9: Hyperparameter Settings for SFT Training Stage.

	Qwen2.5-VL-7B-Instruct-SFT	InternVL2.5-38B-SFT
pretrain model	Qwen2.5-VL-7B-Instruct	InternVL2.5-38B
learning rate	0.5e-5	2e-5
batch size	64	128
optimizer	AdamW	AdamW
lr scheduler	cosine	cosine
image strategy	image_max_pixels=262144	max_dynamic_patch=6
warmup ratio	0.1	0.03
max epochs	1	1
bf16	True	True

Other Details. The training environment consists of CentOS Linux release 7.6.1810 operating system with CUDA 12.1. For Qwen2.5-VL-7B-Instruct-RL, we train for 80 steps on 1×8 A800 GPUs and for InternVL2.5-38B-RL we train for 100 steps on 6×8 A800 GPUs.

C.8 RL MODELS EVALUATION DETAILS

As mentioned above, we apply format rewards in RL experiments. Thus, to fully investigate the models' latent reasoning abilities, we utilize implement training-aligned prompts during evaluation in

Table 10: Hyperparameter Settings for RL Training Stage.

	Qwen2.5-VL-7B-Instruct-RL	InternVL2.5-38B-RL
pretrain model	Qwen2.5-VL-7B-Instruct	InternVL2.5-38B
RL Algorithm	rloo	rloo
train batch size	128	64
rollout batch_size	256	128
temperature	1	1
n samples per prompt	16	8
prompt max len	1024	4096
generate max len	3000	3000
bf16	True	True
actor learning rate	1e-6	1e-6
init kl coef	0	0

VisuLogic, which is shown as follows: "*Solve the complex visual logical reasoning problem through step-by-step reasoning. Think about the reasoning process first and answer the question following this format: <think> THINKING </think><answer> ANSWER </answer>*".

C.9 RESULTS OF RL EXPERIMENTS

As shown in Table 1, MLLMs with reinforcement learning optimization can yield obvious improvements in visual reasoning performance. *Qwen2.5-VL-7B-Instruct-RL* attains 28.0%, a 2.0 percentage point boost over its non-RL counterpart. More strikingly, *InternVL2.5-38B-RL* reaches 31.1%, surpassing the original non-RL model by 5.6% and establishing a new state-of-the-art on VisuLogic. Furthermore, compared to supervised fine-tuning (SFT) on identical datasets, RL-enhanced models demonstrate substantially larger performance gains, underscoring the promise of targeted RL methods for advancing multimodal visual reasoning.

For qualitative analysis of the effectiveness of RL, as illustrated in Figure 7(g), reinforcement learning (RL) encourages deeper, stepwise logical reasoning. The RL-enhanced model successfully captures state transitions (e.g., the movements of chess pieces in Figure 7(a)) and accurately predicts subsequent configurations. Moreover, it learns to iteratively revise intermediate hypotheses—akin to trial-and-error—until a coherent deduction emerges (see additional examples in the Appendix). These findings highlight the potential of RL methods to bolster performance on visual reasoning tasks.

Figures 20, 21, 22, 23, 24 and 25 demonstrate qualitative differences in model outputs between baseline and RL optimized models. It illustrates reinforcement learning (RL) training enables the model to perform fundamental-level analysis of reasoning tasks embedded in graphical representations.

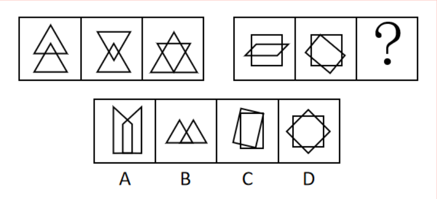


IMAGE DESCRIPTION: The image presents a visual pattern recognition puzzle consisting of a sequence of geometric shapes organized in two rows. The objective is to determine the missing shape in the sequence from the given options.\n\n### Top Row (Sequence of Shapes)\n\n1. **First Shape**: \n - Two identical triangles, one upright and one upside down, overlap each other perfectly to form a star-like shape with a total of six outward-pointing triangular tips.\n - The outlines of the triangles intersect, forming an internal hexagon.\n\n2. **Second Shape**:\n - Two overlapping triangles, one upright and one upside down, slightly offset and not perfectly aligned, creating a symmetric \"X\" formation at the center.\n - The overlap forms a six-pointed star with more distinct and enhanced intersection lines visible inside.\n\n3. **Third Shape**:\n - Two overlapping triangles, one upright and one upside down, arranged symmetrically to create a six-pointed star.\n - The internal lines are more connected, forming a central hexagon similar in style to the first shape, with additional internal lines from the triangle overlaps.\n\n4. **Fourth Shape**:\n - Two overlapping parallelograms, forming a shape that looks like a hexagon outlined in a bold, overlapping manner.\n - The parallelograms intersect to create a pattern with sharp angles and straight lines defining internal quadrilateral shapes.\n\n5. **Fifth Shape**:\n - A centralized octagon with an inner pattern of lines creating a star shape, with the corners of the star connecting each side of the octagon from the center.\n\n6. **Missing Shape (Question Mark)**:\n - This is the space where the next shape in the sequence would logically fit, represented by a question mark.\n\n### Bottom Row (Answer Options)\n\n- **Option A**:\n - Two vertical rectangles with a central pointed triangle merging at the top to form an arrowhead shape.\n - The base of the rectangles is flat, with an internal vertical line separating them.\n\n- **Option B**: \n - Two triangles placed side by side with a small gap between them.\n - The triangles are upright and identical in size and shape.\n\n- **Option C**:\n - Two parallelograms overlapping each other slightly, tilted sideways.\n - The overlapping creates an interesting central intersection with slanting lines.\n\n- **Option D**:\n - An octagon with a central star pattern formed by crossing diagonal lines.\n - The star is symmetrical and fits perfectly within the octagon's outline.\n\n### Task\n\nThe objective is to determine which of the four options (A, B, C, or D) logically completes the sequence in the top row.

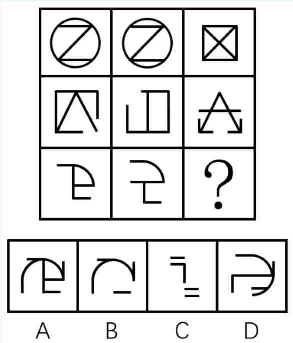


IMAGE DESCRIPTION: The image is a 3x3 grid puzzle with the task of identifying the missing symbol in the bottom right cell. Below the grid are four answer options labeled A, B, C, and D.\n\n### Grid Description\n\n#### First Row:\n1. **Cell 1**: A circle with a diagonal line running from the top left to the bottom right.\n2. **Cell 2**: An identical circle with a diagonal line, from the top left to bottom right.\n3. **Cell 3**: A square with an \"X\" crossing through it.\n\n#### Second Row:\n1. **Cell 4**: A combination of a partial square missing the right side, with a triangle pointing upwards.\n2. **Cell 5**: A partial square missing the left side.\n3. **Cell 6**: An upward-pointing triangle with a horizontal line intersecting it about one-third from the top.\n\n#### Third Row:\n1. **Cell 7**: A mirrored combination of a partial circle on the left and an \"L\" shaped line on the right.\n2. **Cell 8**: Similar to Cell 7, a mirrored partial circle on the left with an extended \"L\" shape on the right.\n3. **Cell 9**: Contains a question mark, indicating the missing symbol in this cell.\n\n### Answer Options\n\n#### Option A:\n- A combination of a mirrored partial circle on the left with a small \"L\" shaped line extending from the right side.\n\n#### Option B:\n- Similar to Option A, but the partial circle has an additional horizontal line.\n\n#### Option C:\n- A vertical and horizontal series of short parallel lines resembling an incomplete rectangle.\n\n#### Option D:\n- An \"L\" shaped line with a mirrored partial circle on the left side, without any additional lines.\n\nThe task involves identifying the pattern or logic that governs the symbols in the grid to select the correct option that completes it

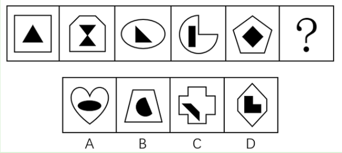


IMAGE DESCRIPTION: The image consists of two rows, each containing boxes with geometric shapes, indicating a puzzle pattern.\n\n### Top Row:\n1. **First Box**: Contains a black equilateral triangle centered within a white square.\n2. **Second Box**: Features two black triangles, one inverted and one upright, forming an hourglass shape, centered inside a pentagon.\n3. **Third Box**: Shows a black triangle pointing right, inside a horizontal white oval.\n4. **Fourth Box**: Displays a black rectangle and a larger black semicircle, creating a shape like the letter \"C,\" inside a circle.\n5. **Fifth Box**: Contains a black diamond inside a pentagon.\n6. **Sixth Box**: Contains a question mark, suggesting the missing piece in the pattern.\n\n### Bottom Row:\nThis row presents four possible answer choices labeled A to D.\n\n- **Option A**: Depicts a horizontal black oval inside a white heart shape.\n- **Option B**: Includes a black shape resembling a curved segment inside a pentagon.\n- **Option C**: Features a black diagonal stripe dissecting a white cross.\n- **Option D**: Displays an L-shaped black figure inside a pentagon.\n\nPatterns in the top row suggest a sequence or rule, and the task is to determine which option (A, B, C, or D) should replace the question mark based on the continuation of the sequence or pattern.

Figure 14: Part of image caption in LLM evaluation.

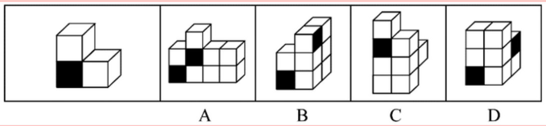


IMAGE DESCRIPTION: The image displays a cube-based puzzle with one reference figure on the left and four options labeled A, B, C, and D to the right. Each option is within its own rectangular boundary.\n\n1. **Reference Figure:**\n - Consists of three cubes configured in an L-shape.\n - The base cube is in the bottom layer.\n - Another cube is placed directly on top of the base cube.\n - The third cube is adjacent to the top cube on the right, forming the L-shape.\n - The front face of the base cube is filled in with black color.\n\n2. **Option A:**\n - Contains six cubes forming a stepped structure.\n - The base consists of two rows, where the left row has two cubes and the right row has three cubes.\n - The top cube of the left row and the middle cube of the right row are stacked on, forming a second layer with two cubes.\n - The left cube in the second layer matches the position of the black face from the reference, so it is filled with black.\n\n3. **Option B:**\n - Composed of six cubes arranged in a vertical stack.\n - The bottom layer has three cubes forming an L-shape.\n - Two cubes are on the left forming a column, with one on the right adjacent to the top of the first cube in the column.\n - The top layer has a single cube directly over the central bottom cube, curving upwards.\n - One of the side faces of the bottom left cube is filled in black.\n\n4. **Option C:**\n - Features five cubes in a half-stair configuration.\n - The configuration starts with two vertical cubes on the left and three cubes forming a step on the right.\n - A vertical column of two cubes stands on the right edge, and another cube is beside the topmost cube of this column.\n - The black face appears on the bottom left cube in a front-facing position.\n\n5. **Option D:**\n - Displays a symmetrical arrangement with a central cube surrounded by four cubes.\n - The configuration resembles a plus sign when viewed from the front.\n - A straightforward 3x3 cube grid with a single black face on the bottom layer, located at the front of the central cube.\n\nEach option offers a different combination of cubes but only one mirrors the black face placement from the initial reference figure.

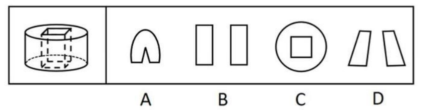


IMAGE DESCRIPTION: The image is a visual puzzle asking which shape of four can be folded to make a given 3D object. It is presented as a single rectangular frame with two sections.\n\n### Left Section:\n- **3D Object**: \n - Depicts a transparent cylinder with a rectangular top and bottom, resembling a drum with squared faces.\n - The cylindrical shape is thin and wide.\n - There is a solid rectangular face shown on top, with dashed lines indicating the back and bottom rectangle edges inside the transparent surface.\n \n### Right Section:\n- **Four 2D Shapes Labeled A to D**:\n \n - **A**: \n - Shape resembling a horseshoe or a U, oriented vertically.\n - Two long, curved edges form the sides connecting a wider curved top and a narrow open bottom.\n\n - **B**: \n - Two identical, separate tall rectangles placed vertically side by side.\n\n - **C**: \n - A circle encompassing a smaller square exactly in the center.\n - The square's sides are parallel to the circle’s diameter, not touching the circle's edge.\n\n - **D**: \n - Two identical trapezoids with longer vertical sides on the left and shorter on the right, slanted inwardly towards each other.\n \nEach shape option is shown as a potential development that could fold into the 3D cylinder on the left.

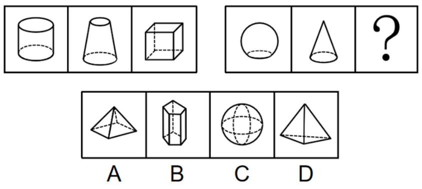
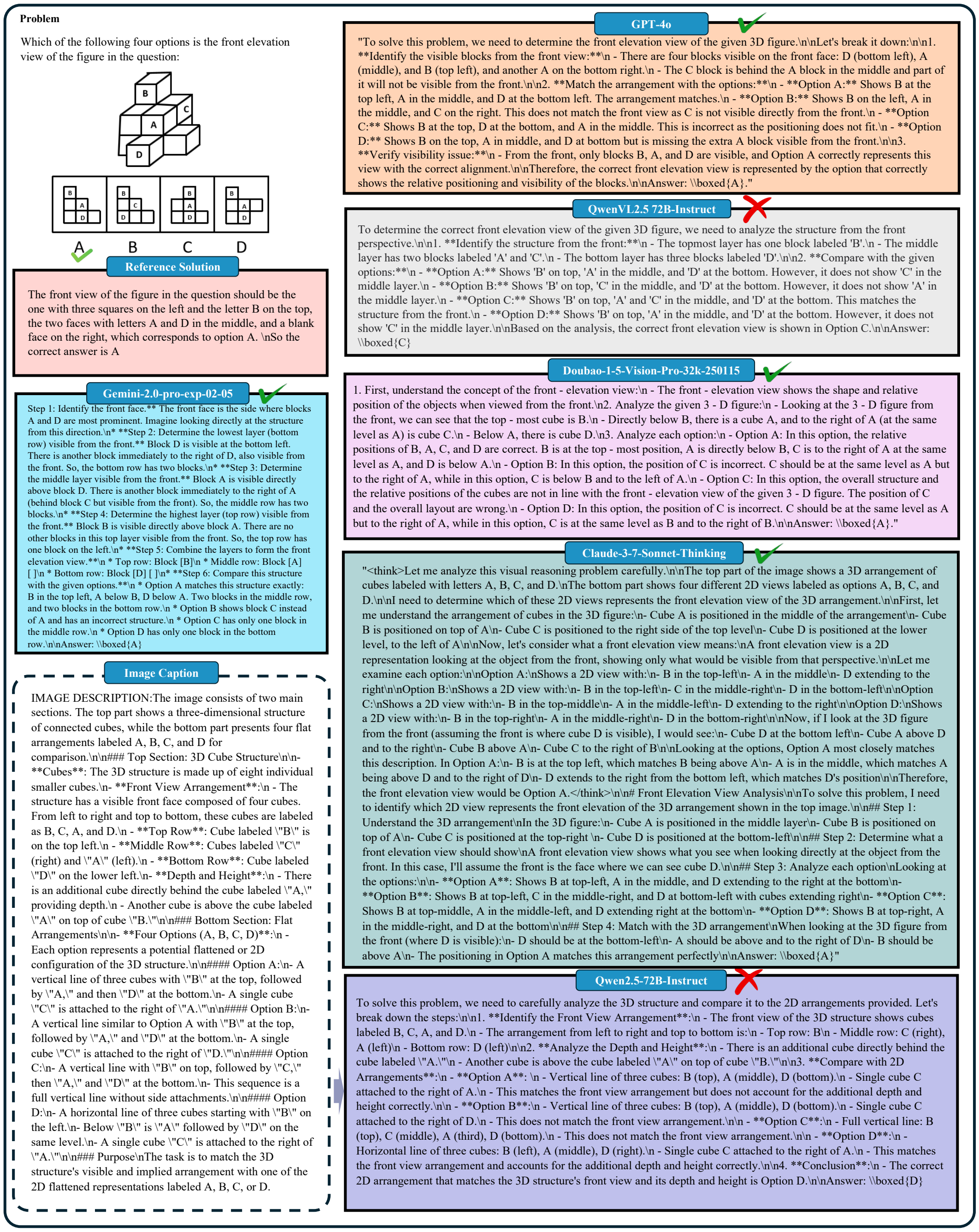


IMAGE DESCRIPTION: The image presents a visual pattern recognition problem involving geometric shapes, divided into two main parts: the problem set on top and four possible answer choices below.\n\n### Problem Set (Top Part):\n- **First Row (Left to Right):**\n 1. **Cylinder:** A three-dimensional shape with two parallel circular bases connected by a curved surface, depicted with dashed lines indicating hidden edges.\n 2. **Truncated Cone (Frustum):** A shape with a circular base and a smaller circular top, connected by a curved surface. The top is parallel and smaller than the base, also shown with dashed lines for hidden edges.\n 3. **Cube:** A solid shape with six equal square faces. The hidden edges are represented with dashed lines to suggest depth.\n\n- **Second Row (Left to Right):**\n 1. **Circle:** A flat, two-dimensional shape with only the perimeter shown.\n 2. **Cone:** A three-dimensional shape with a circular base tapering to a point (apex) at the top, depicted with dashed lines to show hidden parts of the base.\n 3. **Question Mark:** A placeholder indicating the missing shape, suggesting the need to find a shape that logically follows the presented sequence.\n\n### Answer Choices (Bottom Part):\nEach option is labeled with a capital letter below each shape from A to D.\n\n- **Option A:**\n - **Square Pyramid:** A three-dimensional shape with a square base and four triangular faces that meet at a single apex. Dashed lines indicate the hidden part of the base.\n \n- **Option B:**\n - **Hexagonal Prism:** A solid with two parallel hexagonal faces and rectangular faces connecting them. Dashed lines represent the hidden edges of the hexagons.\n \n- **Option C:**\n - **Sphere:** A perfect three-dimensional round shape. Equator and meridian lines are shown in dashed format to suggest depth and roundness.\n\n- **Option D:**\n - **Triangular Pyramid (Tetrahedron):** A three-dimensional shape with four triangular faces, one of which is the base. Dashed lines indicate the hidden edges.\n\n### Task:\nThe task is to identify the shape from the answer choices (A, B, C, D) that logically completes the sequence in the problem set, based on the pattern from the first row where three-dimensional shapes were involved, and consistent with the existing sequence pattern.

Figure 15: Part of image caption in LLM evaluation.



Problem

Which of the following four options is the front elevation view of the figure in the question:

Reference Solution

The front view of the figure in the question should be the one with three squares on the left and the letter B on the top, the two faces with letters A and D in the middle, and a blank face on the right, which corresponds to option A. \nSo the correct answer is A

Gemini-2.0-pro-exp-02-05

Step 1: Identify the front face.** The front face is the side where blocks A and D are most prominent. Imagine looking directly at the structure from this direction.\n* **Step 2: Determine the lowest layer (bottom row) visible from the front.** Block D is visible at the bottom left. There is another block immediately to the right of D, also visible from the front. So, the bottom row has two blocks.\n* **Step 3: Determine the middle layer visible from the front.** Block A is visible directly above block D. There is another block immediately to the right of A (behind block C but visible from the front). So, the middle row has two blocks.\n* **Step 4: Determine the highest layer (top row) visible from the front.** Block B is visible directly above block A. There are no other blocks in this top layer visible from the front. So, the top row has one block on the left.\n* **Step 5: Combine the layers to form the front elevation view.**\n * Top row: Block [B]\n * Middle row: Block [A] [ ]\n * Bottom row: Block [D] [ ]\n* **Step 6: Compare this structure with the given options.**\n * Option A matches this structure exactly: B in the top left, A below B, D below A. Two blocks in the middle row, and two blocks in the bottom row.\n * Option B shows block C instead of A and has an incorrect structure.\n * Option C has only one block in the middle row.\n * Option D has only one block in the bottom row.\n\nAnswer: \\boxed{A}

Image Caption

IMAGE DESCRIPTION:The image consists of two main sections. The top part shows a three-dimensional structure of connected cubes, while the bottom part presents four flat arrangements labeled A, B, C, and D for comparison.\n\n### Top Section: 3D Cube Structure\n\n- **Cubes**: The 3D structure is made up of eight individual smaller cubes.\n- **Front View Arrangement**:\n - The structure has a visible front face composed of four cubes. From left to right and top to bottom, these cubes are labeled as B, C, A, and D.\n - **Top Row**: Cube labeled \"B\" is on the top left.\n - **Middle Row**: Cubes labeled \"C\" (right) and \"A\" (left).\n - **Bottom Row**: Cube labeled \"D\" on the lower left.\n- **Depth and Height**:\n - There is an additional cube directly behind the cube labeled \"A,\" providing depth.\n - Another cube is above the cube labeled \"A\" on top of cube \"B.\"\n\n### Bottom Section: Flat Arrangements\n\n- **Four Options (A, B, C, D)**:\n - Each option represents a potential flattened or 2D configuration of the 3D structure.\n\n#### Option A:\n- A vertical line of three cubes with \"B\" at the top, followed by \"A,\" and then \"D\" at the bottom.\n- A single cube \"C\" is attached to the right of \"A.\"\n\n#### Option B:\n- A vertical line similar to Option A with \"B\" at the top, followed by \"A,\" and \"D\" at the bottom.\n- A single cube \"C\" is attached to the right of \"D.\"\n\n#### Option C:\n- A vertical line with \"B\" on top, followed by \"C,\" then \"A,\" and \"D\" at the bottom.\n- This sequence is a full vertical line without side attachments.\n\n#### Option D:\n- A horizontal line of three cubes starting with \"B\" on the left.\n- Below \"B\" is \"A\" followed by \"D\" on the same level.\n- A single cube \"C\" is attached to the right of \"A.\"\n\n### Purpose\nThe task is to match the 3D structure's visible and implied arrangement with one of the 2D flattened representations labeled A, B, C, or D.

GPT-4o

"To solve this problem, we need to determine the front elevation view of the given 3D figure.\n\nLet's break it down:\n\n1. **Identify the visible blocks from the front view:**\n - There are four blocks visible on the front face: D (bottom left), A (middle), and B (top left), and another A on the bottom right.\n - The C block is behind the A block in the middle and part of it will not be visible from the front.\n\n2. **Match the arrangement with the options:**\n - **Option A:** Shows B at the top left, A in the middle, and D at the bottom left. The arrangement matches.\n - **Option B:** Shows B on the left, A in the middle, and C on the right. This does not match the front view as C is not visible directly from the front.\n - **Option C:** Shows B at the top, D at the bottom, and A in the middle. This is incorrect as the positioning does not fit.\n - **Option D:** Shows B on the top, A in middle, and D at bottom but is missing the extra A block visible from the front.\n\n3. **Verify visibility issue:**\n - From the front, only blocks B, A, and D are visible, and Option A correctly represents this view with the correct alignment.\n\nTherefore, the correct front elevation view is represented by the option that correctly shows the relative positioning and visibility of the blocks.\n\nAnswer: \\boxed{A}."

QwenVL2.5 72B-Instruct

To determine the correct front elevation view of the given 3D figure, we need to analyze the structure from the front perspective.\n\n1. **Identify the structure from the front:**\n - The topmost layer has one block labeled 'B'.\n - The middle layer has two blocks labeled 'A' and 'C'.\n - The bottom layer has three blocks labeled 'D'.\n\n2. **Compare with the given options:**\n - **Option A:** Shows 'B' on top, 'A' in the middle, and 'D' at the bottom. However, it does not show 'C' in the middle layer.\n - **Option B:** Shows 'B' on top, 'C' in the middle, and 'D' at the bottom. However, it does not show 'A' in the middle layer.\n - **Option C:** Shows 'B' on top, 'A' and 'C' in the middle, and 'D' at the bottom. This matches the structure from the front.\n - **Option D:** Shows 'B' on top, 'A' in the middle, and 'D' at the bottom. However, it does not show 'C' in the middle layer.\n\nBased on the analysis, the correct front elevation view is shown in Option C.\n\nAnswer: \\boxed{C}

Doubao-1-5-Vision-Pro-32k-250115

1. First, understand the concept of the front - elevation view:\n - The front - elevation view shows the shape and relative position of the objects when viewed from the front.\n2. Analyze the given 3 - D figure:\n - Looking at the 3 - D figure from the front, we can see that the top - most cube is B.\n - Directly below B, there is a cube A, and to the right of A (at the same level as A) is cube C.\n - Below A, there is cube D.\n3. Analyze each option:\n - Option A: In this option, the relative positions of B, A, C, and D are correct. B is at the top - most position, A is directly below B, C is to the right of A at the same level as A, and D is below A.\n - Option B: In this option, the position of C is incorrect. C should be at the same level as A but to the right of A, while in this option, C is below B and to the left of A.\n - Option C: In this option, the overall structure and the relative positions of the cubes are not in line with the front - elevation view of the given 3 - D figure. The position of C and the overall layout are wrong.\n - Option D: In this option, the position of C is incorrect. C should be at the same level as A but to the right of A, while in this option, C is at the same level as B and to the right of B.\n\nAnswer: \\boxed{A}."

Claude-3-7-Sonnet-Thinking

"<think>Let me analyze this visual reasoning problem carefully.\n\nThe top part of the image shows a 3D arrangement of cubes labeled with letters A, B, C, and D.\nThe bottom part shows four different 2D views labeled as options A, B, C, and D.\n\nI need to determine which of these 2D views represents the front elevation view of the 3D arrangement.\n\nFirst, let me understand the arrangement of cubes in the 3D figure:\n- Cube A is positioned in the middle of the arrangement\n- Cube B is positioned on top of A\n- Cube C is positioned to the right side of the top level\n- Cube D is positioned at the lower level, to the left of A\n\nNow, let's consider what a front elevation view means:\nA front elevation view is a 2D representation looking at the object from the front, showing only what would be visible from that perspective.\n\nLet me examine each option:\n\nOption A:\nShows a 2D view with:\n- B in the top-left\n- A in the middle\n- D extending to the right\n\nOption B:\nShows a 2D view with:\n- B in the top-left\n- C in the middle-right\n- D in the bottom-left\n\nOption C:\nShows a 2D view with:\n- B in the top-middle\n- A in the middle-left\n- D extending to the right\n\nOption D:\nShows a 2D view with:\n- B in the top-right\n- A in the middle-right\n- D in the bottom-right\n\nNow, if I look at the 3D figure from the front (assuming the front is where cube D is visible), I would see:\n- Cube D at the bottom left\n- Cube A above D and to the right\n- Cube B above A\n- Cube C to the right of B\n\nLooking at the options, Option A most closely matches this description. In Option A:\n- B is at the top left, which matches B being above A\n- A is in the middle, which matches A being above D and to the right of D\n- D extends to the right from the bottom left, which matches D's position\n\nTherefore, the front elevation view would be Option A.</think>\n\n# Front Elevation View Analysis\n\nTo solve this problem, I need to identify which 2D view represents the front elevation of the 3D arrangement shown in the top image.\n\n## Step 1: Understand the 3D arrangement\nIn the 3D figure:\n- Cube A is positioned in the middle layer\n- Cube B is positioned on top of A\n- Cube C is positioned at the top-right \n- Cube D is positioned at the bottom-left\n\n## Step 2: Determine what a front elevation view should show\nA front elevation view shows what you see when looking directly at the object from the front. In this case, I'll assume the front is the face where we can see cube D.\n\n## Step 3: Analyze each option\nLooking at the options:\n\n- **Option A**: Shows B at top-left, A in the middle, and D extending to the right at the bottom\n- **Option B**: Shows B at top-left, C in the middle-right, and D at bottom-left with cubes extending right\n- **Option C**: Shows B at top-middle, A in the middle-left, and D extending right at the bottom\n- **Option D**: Shows B at top-right, A in the middle-right, and D at the bottom\n\n## Step 4: Match with the 3D arrangement\nWhen looking at the 3D figure from the front (where D is visible):\n- D should be at the bottom-left\n- A should be above and to the right of D\n- B should be above A\n- The positioning in Option A matches this arrangement perfectly\n\nAnswer: \\boxed{A}"

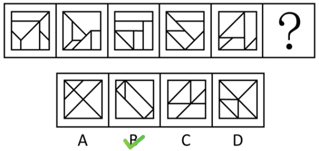
Qwen2.5-72B-Instruct

To solve this problem, we need to carefully analyze the 3D structure and compare it to the 2D arrangements provided. Let's break down the steps:\n\n1. **Identify the Front View Arrangement**:\n - The front view of the 3D structure shows cubes labeled B, C, A, and D.\n - The arrangement from left to right and top to bottom is:\n - Top row: B\n - Middle row: C (right), A (left)\n - Bottom row: D (left)\n\n2. **Analyze the Depth and Height**:\n - There is an additional cube directly behind the cube labeled \"A.\"\n - Another cube is above the cube labeled \"A\" on top of cube \"B.\"\n\n3. **Compare with 2D Arrangements**:\n - **Option A**: \n - Vertical line of three cubes: B (top), A (middle), D (bottom).\n - Single cube C attached to the right of A.\n - This matches the front view arrangement but does not account for the additional depth and height correctly.\n\n - **Option B**:\n - Vertical line of three cubes: B (top), A (middle), D (bottom).\n - Single cube C attached to the right of D.\n - This does not match the front view arrangement.\n\n - **Option C**:\n - Full vertical line: B (top), C (middle), A (third), D (bottom).\n - This does not match the front view arrangement.\n\n - **Option D**:\n - Horizontal line of three cubes: B (left), A (middle), D (right).\n - Single cube C attached to the right of A.\n - This matches the front view arrangement and accounts for the additional depth and height correctly.\n\n4. **Conclusion**:\n - The correct 2D arrangement that matches the 3D structure's front view and its depth and height is Option D.\n\nAnswer: \\boxed{D}

Figure 16: **Solution examples generated by different models.** Reference solution and outputs generated by GPT-4o, Qwen2.5VL-72B-Instruct, Gemini-2.0-pro-exp-02-05, Doubao-1.5-Vision-Pro-32K and Claude-3.7-sonnet-thinking. Additionally, the image caption and solution from LLMs (Qwen2.5-72B-Instruct) are also illustrated.

Problem

Which of the following four options is the front elevation view of the figure in the question:



Reference Solution

It is found that the closed area of the figure in the question stem is obvious. Considering the number of faces, the number of faces in the figure in the question stem and the option figures is 6, so it is impossible to choose the answer. Considering the refinement of the faces, it is found again that the figure in the question stem basically contains triangular faces. Considering the number of triangular faces, the number of triangular faces in the figure in the question stem is 0, 1, 2, 3, and 4 in sequence, so the number of triangular faces in the figure at "?" should be 5, and only option B meets the requirement. \nSo the correct answer is B.

Kimi-latest ✓

To solve this problem, we need to identify the pattern in the sequence of the first five figures and then determine which of the options (A, B, C, or D) fits this pattern.\n\n1. **Observe the Sequence**: Look at the sequence of the first five figures. Notice that each figure is a square divided into four smaller triangles by two diagonals that intersect at the center of the square.\n\n2. **Identify the Pattern**: Observe how the triangles are shaded in each figure:\n - The first figure has the top-left and bottom-right triangles shaded.\n - The second figure has the top-right and bottom-left triangles shaded.\n - The third figure has the top-left and bottom-right triangles shaded again.\n - The fourth figure has the top-right and bottom-left triangles shaded again.\n - The fifth figure has the top-left and bottom-right triangles shaded.\n\n3. **Determine the Pattern**: The pattern alternates between shading the top-left and bottom-right triangles and shading the top-right and bottom-left triangles.\n\n4. **Apply the Pattern**: Since the fifth figure has the top-left and bottom-right triangles shaded, the next figure in the sequence should have the top-right and bottom-left triangles shaded.\n\n5. **Match with Options**: Look at the options (A, B, C, D) and identify which one matches the pattern of having the top-right and bottom-left triangles shaded.\n - Option A: Top-left and bottom-right triangles shaded.\n - Option B: Top-right and bottom-left triangles shaded.\n - Option C: Top-left and bottom-right triangles shaded.\n - Option D: Top-right and bottom-left triangles shaded.\n\n6. **Select the Correct Option**: Both options B and D have the top-right and bottom-left triangles shaded. However, we need to ensure that the pattern is consistent with the previous figures. Since the sequence alternates and the fifth figure has the top-left and bottom-right triangles shaded, the next figure should indeed have the top-right and bottom-left triangles shaded. Both B and D fit this pattern, but we need to choose one. Given the options, we can select either B or D. For the sake of this solution, we will choose option B.\n\nTherefore, the most suitable option to fill in the question mark is \\boxed{B}.

Image Caption

IMAGE DESCRIPTION:The image presents a puzzle problem made up of a series of geometric patterns. The top row consists of five squares in a sequence, followed by a question mark indicating a missing pattern in a sixth square. Below this row, there are four options labeled A, B, C, and D, each containing a unique geometric pattern inside a square.\n\n### Top Row (Pattern Series):\n\n1. **Square 1**: \n - Contains a large diagonal line from the bottom left to the top right, dividing the square into a triangular shape on the bottom left.\n - Within the upper right section, a vertical and horizontal line meet, forming smaller triangles and a small trapezoid.\n\n2. **Square 2**: \n - The pattern has a large diagonal from the top left to the bottom right.\n - An additional horizontal line creates two triangles on either side with a small square or rectangle in the center.\n\n3. **Square 3**: \n - A series of horizontal and vertical lines divide the square into rectangles and a smaller square.\n - There is also a diagonal line connecting two non-adjacent corners, intersecting some of these smaller shapes.\n\n4. **Square 4**: \n - Features a central diagonal line from the top right to the bottom left, creating two large triangles.\n - Additional lines partition these triangles into smaller triangles and trapezoids.\n\n5. **Square 5**: \n - Contains a vertical line dividing the square in half, with a horizontal line at the mid-top extending partially.\n - A diagonal runs from the top left to the bottom right corner through these lines, creating several intersecting triangular regions.\n\n6. **Question Mark**: \n - A placeholder representing the missing sixth square pattern that needs to be identified.\n\n### Bottom Row (Options A, B, C, D):\n\n- **Option A**: \n - The pattern has multiple intersecting diagonal lines creating a star-like effect with eight triangle segments.\n\n- **Option B**: \n - Features a large diagonal spanning from the top left to the bottom right.\n - Additional lines form a mix of triangles and quadrilateral shapes, leaning towards asymmetry.\n\n- **Option C**: \n - Contains intersecting vertical, horizontal, and diagonal lines.\n - Several small triangles and quadrilaterals are evident, creating a structured yet fragmented pattern.\n\n- **Option D**: \n - Similar to Option A, multiple intersecting diagonals create multiple triangular segments.\n - The design is symmetric and balanced.\n\n### Task:\nThe task is to determine which of the four options (A, B, C, D) logically continues the sequence established by the first five squares in the top row. \n\nEach square and option is enclosed with a solid border, emphasizing the contained geometric pattern. The solution involves recognizing a pattern or rule that connects the top sequence with one of the provided options.

GPT-4o ✓

To solve the problem, let's analyze the sequence of figures:\n\n1. **Identify Patterns in Given Figures:**\n - Break down each figure in the sequence to identify patterns or transformations.\n - The first image contains a large square divided into several triangular sections.\n - The second image shows an edge merging and forming a new shape.\n - The third image introduces more fragmented parts.\n - The fourth image repeats a similar pattern with a new twist in fragmentation.\n2. **Look at the Specific Transformations:**\n - Notice possible rotations, reflections, or consistent transformations that occur from one figure to the next.\n - There is a pattern of dividing the square and incrementing triangular or rectangular partitions in each step.\n\n3. **Predict the Next Transformation:**\n - By examining figures, notice that each set seems to introduce a new line or rotate slightly and changes align with a simple geometric transformation.\n - The consistent theme of geometric division is essential.\n\n4. **Applying the Observed Pattern:**\n - Look at options A, B, C, and D.\n - Option A features a very intricate pattern with no straightforward changes observed in previous steps.\n - Option B shows a diagonal shape consistent with transformations.\n - Option C and D suggest other structural changes that fit less with the observed pattern.\n - Focus on selection that continues simple progression of incremental shape division or mirror.\n\n5. **Select the Answer**\n - Consider all evidence and the goal for minimal transformation matching observed traits.\n\nAnswer: \\boxed{B}

Doubao-1-5-Vision-Pro-32k-250115 ✓

1. First, analyze the characteristics of the figures in the question - set:\n - Each figure in the question - set is composed of a large square divided into several small right - angled triangles. We can consider the relationship between the lines in the figures, such as the parallel and perpendicular relationships of the line segments.\n - Observe the number of right - angled triangles in each figure: The first figure has 5 right - angled triangles, the second figure has 4 right - angled triangles, the third figure has 3 right - angled triangles, the fourth figure has 2 right - angled triangles, and the fifth figure has 1 right - angled triangle. So, the number of right - angled triangles in the figures in the question - set is decreasing by 1 successively.\n2. Then, analyze the options:\n - Option A: The figure has 2 right - angled triangles, which does not meet the requirement of the number of right - angled triangles decreasing by 1 successively.\n - Option B: The figure has 0 right - angled triangles. According to the rule that the number of right - angled triangles in the figures in the question - set is 5, 4, 3, 2, 1, and then 0, this option meets the rule.\n - Option C: The figure has 4 right - angled triangles, which does not meet the rule.\n - Option D: The figure has 1 right - angled triangle, which does not meet the rule.\n\nAnswer: \\boxed{B}.

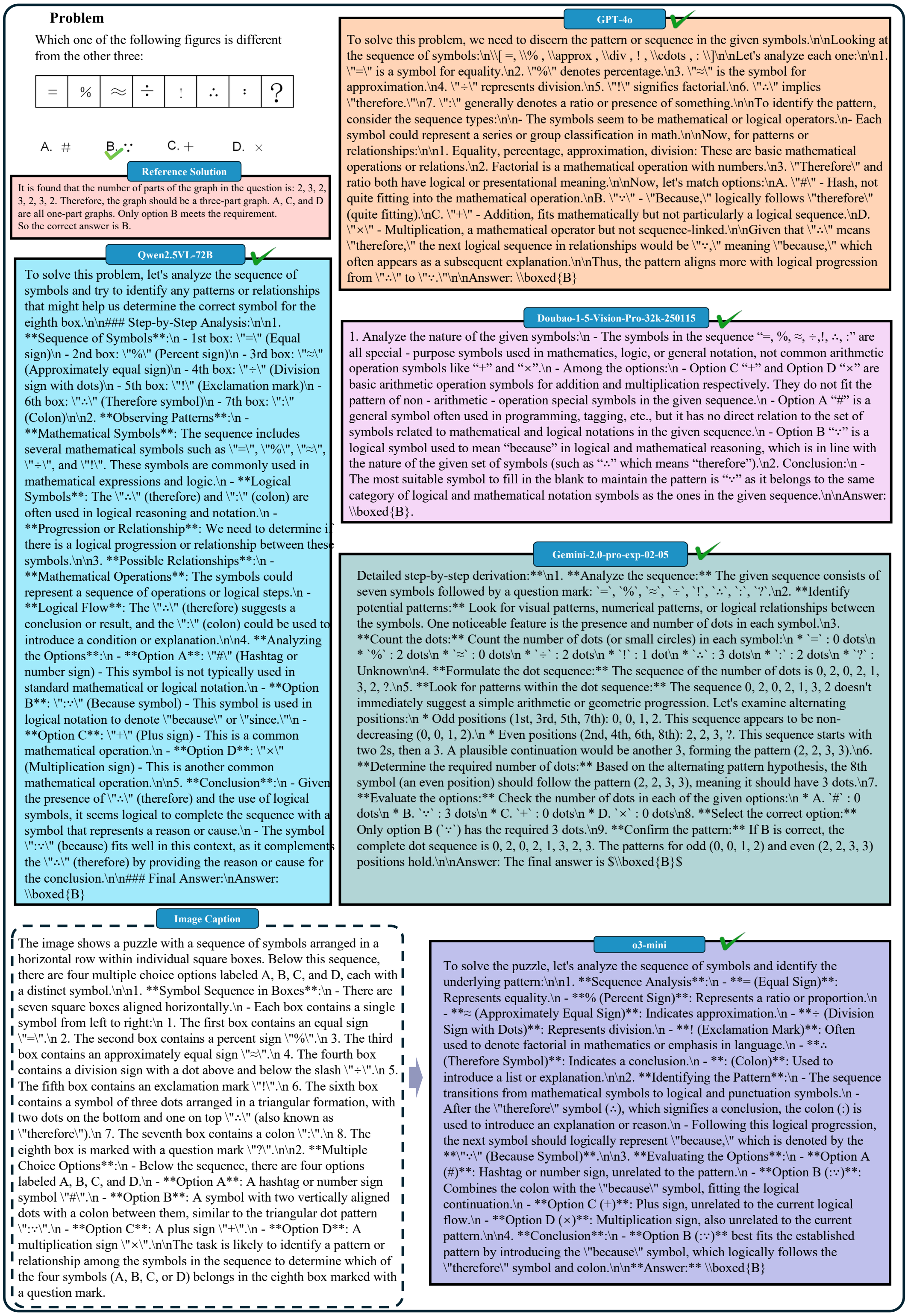
Gemini-2.0-pro-exp-02-05 ✗

Analyze the sequence: Observe the five figures provided in the sequence (let's call them F1, F2, F3, F4, F5) and look for a pattern or relationship between them to determine the sixth figure (F6).\n* **Examine structural features:** Count line segments, analyze shapes, check symmetry, look at intersections, etc. Each figure consistently has 6 internal line segments. The shapes and symmetries vary significantly.\n* **Look for simple transformations:** Rotations or reflections do not seem to map consecutive images directly (e.g., F1 rotated/reflected is not F2).\n* **Consider alternating patterns or cycles:**\n * Maybe odd-numbered figures (F1, F3, F5) follow one pattern, and even-numbered figures (F2, F4, F6) follow another.\n * Maybe there's a cycle of a specific length (e.g., 4). If the cycle is 4, then F5 should relate to F1, and F6 should relate to F2. Visually, F5 is quite different from F1.\n* **Hypothesize a relationship with a fixed offset:** Let's test if there's a consistent transformation T such that F(n+k) = T(F(n)) for some k. Let's try k=3.\n * **Check F1 -> F4:**\n * F1 has a central structure based on connecting midpoints (a diamond shape) plus vertical line segments connecting the diamond to the center point (or near it). Let's denote this structure as (MD + |).\n * F4 has a central structure based on the two main diagonals (an 'X' shape) plus two smaller diagonal segments (slope -1) in the bottom-left and top-right corners relative to the center. Let's denote this structure as (X + \\_shifted).\n * The transformation T1->4 changes the base structure (MD -> X) and the added lines (| -> \\_shifted).\n * **Check F2 -> F5:**\n * F2 has a complex structure.\n * F5 has one main diagonal ('/' shape) plus three other line segments. Let's denote this as (/ + L).\n * The transformation T2->5 changes the base structure (F2 -> /) and the added lines (F2 -> L). This fits the general idea of \"Change Base, Change Additions\".\n * **Predict F3 -> F6:**\n * F3 has a central structure based on horizontal and vertical midlines (a '+' shape) plus two diagonal segments (slope -1) in the top-left and bottom-right corners relative to the center. Let's denote this structure as (+ + \\).\n * Apply the transformation logic: Change Base, Change Additions. Following the pattern seen in T1->4 and potentially T3->F6 (where the base also becomes X), let's assume the base (+) changes to (X). What happens to the added lines (\\ diagonals)? In T1->4, | changed to \\. In T3->F6, maybe \\ changes to |?\n * If (+) changes to (X) and (\\) changes to (|), the resulting figure F6 would be (X + |). This means an 'X' shape from the main diagonals plus vertical line segments. Let's check the options.\n * Option A: (X + -) - Diagonal cross plus horizontal connectors.\n * Option B: (/ + complex) - Slash diagonal plus other lines.\n * Option C: (/ + complex) - Slash diagonal plus other lines.\n * Option D: (X + |) - Diagonal cross plus vertical connectors.\n * The predicted structure (X + |) matches Option D exactly.\n\n* **Conclusion:** The pattern F(n+3) = Transform(F(n)) holds consistently, where the transformation involves changing the base structure and the type/position of added lines. Applying this transformation to F3 suggests F6 should be Option D.\n\nAnswer: \\boxed{D}

Qwen2.5-72B-Instruct ✓

To solve this visual logical reasoning problem, let's analyze the patterns and transformations occurring from one square to the next in the top row. We will look for consistent changes or rules that can help us predict the next pattern.\n\n### Step-by-Step Analysis:\n\n1. **Square 1 to Square 2**:\n - **Initial Pattern**: Large diagonal from bottom left to top right.\n - **Transformation**: The diagonal flips to go from top left to bottom right, and an additional horizontal line is added, creating more segments.\n\n2. **Square 2 to Square 3**:\n - **Initial Pattern**: Large diagonal from top left to bottom right with a horizontal line.\n - **Transformation**: Additional vertical and horizontal lines are added, creating a grid-like structure with a diagonal intersecting it.\n\n3. **Square 3 to Square 4**:\n - **Initial Pattern**: Grid-like structure with a diagonal.\n - **Transformation**: The grid simplifies to a central diagonal from top right to bottom left, with additional lines creating smaller triangles and trapezoids.\n\n4. **Square 4 to Square 5**:\n - **Initial Pattern**: Central diagonal from top right to bottom left with additional lines.\n - **Transformation**: The pattern simplifies further with a vertical line dividing the square in half and a horizontal line at the mid-top, with a diagonal running through them.\n\n### Observations:\n- **Diagonal Flipping**: The direction of the main diagonal alternates between squares.\n- **Line Addition and Simplification**: Additional lines are added to create more segments, but then the pattern simplifies in the next square.\n- **Symmetry and Complexity**: The complexity of the pattern increases and then decreases, suggesting a cyclical nature.\n\n### Predicting the Next Pattern:\nGiven the alternating diagonal and the cyclical nature of complexity, the next pattern should:\n- Have a diagonal from bottom left to top right (flipping back).\n- Be relatively simple, possibly with fewer additional lines.\n\n### Evaluating the Options:\n- **Option A**: Multiple intersecting diagonals creating a star-like effect. This is too complex and does not follow the simplicity trend.\n- **Option B**: Large diagonal from top left to bottom right with additional lines forming a mix of shapes. This is a good candidate as it follows the diagonal flip and has a moderate level of complexity.\n- **Option C**: Intersecting vertical, horizontal, and diagonal lines. This is also complex and does not fit the simplification trend.\n- **Option D**: Multiple intersecting diagonals creating multiple triangular segments. This is similar to Option A and is too complex.\n\n### Conclusion:\nOption B is the most suitable choice as it follows the diagonal flip and maintains a moderate level of complexity, fitting the observed pattern.\n\nAnswer: \\boxed{B}

Figure 17: **Solution examples generated by different models.** Reference solution and outputs generated by GPT-4o, Kimi-latest, Gemini-2.0-pro-exp-02-05 and Doubao-1.5-Vision-Pro-32K. Additionally, the image caption and solution from LLMs (Qwen2.5-72B-Instruct) are also illustrated.



Problem

Which one of the following figures is different from the other three:

=	%	≈	÷	!	∴	:	?

A. # [✓ B. ∵] C. + D. ×

Reference Solution

It is found that the number of parts of the graph in the question is: 2, 3, 2, 3, 2, 3, 2. Therefore, the graph should be a three-part graph. A, C, and D are all one-part graphs. Only option B meets the requirement.
So the correct answer is B.

Qwen2.5VL-72B ✓

To solve this problem, let's analyze the sequence of symbols and try to identify any patterns or relationships that might help us determine the correct symbol for the eighth box.\n\n### Step-by-Step Analysis:\n\n1. **Sequence of Symbols**:\n - 1st box: \"=\" (Equal sign)\n - 2nd box: \"%\" (Percent sign)\n - 3rd box: \"≈\" (Approximately equal sign)\n - 4th box: \"÷\" (Division sign with dots)\n - 5th box: \"!\" (Exclamation mark)\n - 6th box: \"∴\" (Therefore symbol)\n - 7th box: \":\" (Colon)\n\n2. **Observing Patterns**:\n - **Mathematical Symbols**: The sequence includes several mathematical symbols such as \"=\", \"%\", \"≈\", \"÷\", and \"!\". These symbols are commonly used in mathematical expressions and logic.\n - **Logical Symbols**: The \"∴\" (therefore) and \":\" (colon) are often used in logical reasoning and notation.\n - **Progression or Relationship**: We need to determine if there is a logical progression or relationship between these symbols.\n\n3. **Possible Relationships**:\n - **Mathematical Operations**: The symbols could represent a sequence of operations or logical steps.\n - **Logical Flow**: The \"∴\" (therefore) suggests a conclusion or result, and the \":\" (colon) could be used to introduce a condition or explanation.\n\n4. **Analyzing the Options**:\n - **Option A**: \"#\" (Hashtag or number sign) - This symbol is not typically used in standard mathematical or logical notation.\n - **Option B**: \":∵\" (Because symbol) - This symbol is used in logical notation to denote \"because\" or \"since.\"\n - **Option C**: \"+\" (Plus sign) - This is a common mathematical operation.\n - **Option D**: \"×\" (Multiplication sign) - This is another common mathematical operation.\n\n5. **Conclusion**:\n - Given the presence of \"∴\" (therefore) and the use of logical symbols, it seems logical to complete the sequence with a symbol that represents a reason or cause.\n - The symbol \":∵\" (because) fits well in this context, as it complements the \"∴\" (therefore) by providing the reason or cause for the conclusion.\n\n### Final Answer:\nAnswer: \\boxed{B}

Image Caption

The image shows a puzzle with a sequence of symbols arranged in a horizontal row within individual square boxes. Below this sequence, there are four multiple choice options labeled A, B, C, and D, each with a distinct symbol.\n\n1. **Symbol Sequence in Boxes**:\n - There are seven square boxes aligned horizontally.\n - Each box contains a single symbol from left to right:\n 1. The first box contains an equal sign \"=\".\n 2. The second box contains a percent sign \"%\".\n 3. The third box contains an approximately equal sign \"≈\".\n 4. The fourth box contains a division sign with a dot above and below the slash \"÷\".\n 5. The fifth box contains an exclamation mark \"!\".\n 6. The sixth box contains a symbol of three dots arranged in a triangular formation, with two dots on the bottom and one on top \"∴\" (also known as \"therefore\").\n 7. The seventh box contains a colon \":\".\n 8. The eighth box is marked with a question mark \"?\".\n\n2. **Multiple Choice Options**:\n - Below the sequence, there are four options labeled A, B, C, and D.\n - **Option A**: A hashtag or number sign symbol \"#\".\n - **Option B**: A symbol with two vertically aligned dots with a colon between them, similar to the triangular dot pattern \":∵\".\n - **Option C**: A plus sign \"+\".\n - **Option D**: A multiplication sign \"×\".\n\nThe task is likely to identify a pattern or relationship among the symbols in the sequence to determine which of the four symbols (A, B, C, or D) belongs in the eighth box marked with a question mark.

GPT-4o ✓

To solve this problem, we need to discern the pattern or sequence in the given symbols.\n\nLooking at the sequence of symbols:\n\\[=, \\% , \\approx , \\div , ! , \\cdots , : \\]\n\nLet's analyze each one:\n\n1. \"=\" is a symbol for equality.\n2. \"%\" denotes percentage.\n3. \"≈\" is the symbol for approximation.\n4. \"÷\" represents division.\n5. \"!\" signifies factorial.\n6. \"∴\" implies \"therefore.\"\n7. \":\" generally denotes a ratio or presence of something.\n\nTo identify the pattern, consider the sequence types:\n\n- The symbols seem to be mathematical or logical operators.\n- Each symbol could represent a series or group classification in math.\n\nNow, for patterns or relationships:\n\n1. Equality, percentage, approximation, division: These are basic mathematical operations or relations.\n2. Factorial is a mathematical operation with numbers.\n3. \"Therefore\" and ratio both have logical or presentational meaning.\n\nNow, let's match options:\nA. \"#\" - Hash, not quite fitting into the mathematical operation.\nB. \"∵\" - \"Because,\" logically follows \"therefore\" (quite fitting).\nC. \"+\" - Addition, fits mathematically but not particularly a logical sequence.\nD. \"×\" - Multiplication, a mathematical operator but not sequence-linked.\n\nGiven that \"∴\" means \"therefore,\" the next logical sequence in relationships would be \"∵,\" meaning \"because,\" which often appears as a subsequent explanation.\n\nThus, the pattern aligns more with logical progression from \"∴\" to \"∵.\"\n\nAnswer: \\boxed{B}

Doubao-1-5-Vision-Pro-32k-250115 ✓

1. Analyze the nature of the given symbols:\n - The symbols in the sequence “=, %, ≈, ÷,!, ∴, :” are all special - purpose symbols used in mathematics, logic, or general notation, not common arithmetic operation symbols like “+” and “×”.\n - Among the options:\n - Option C “+” and Option D “×” are basic arithmetic operation symbols for addition and multiplication respectively. They do not fit the pattern of non - arithmetic - operation special symbols in the given sequence.\n - Option A “#” is a general symbol often used in programming, tagging, etc., but it has no direct relation to the set of symbols related to mathematical and logical notations in the given sequence.\n - Option B “∵” is a logical symbol used to mean “because” in logical and mathematical reasoning, which is in line with the nature of the given set of symbols (such as “∴” which means “therefore”).\n2. Conclusion:\n - The most suitable symbol to fill in the blank to maintain the pattern is “∵” as it belongs to the same category of logical and mathematical notation symbols as the ones in the given sequence.\n\nAnswer: \\boxed{B}.

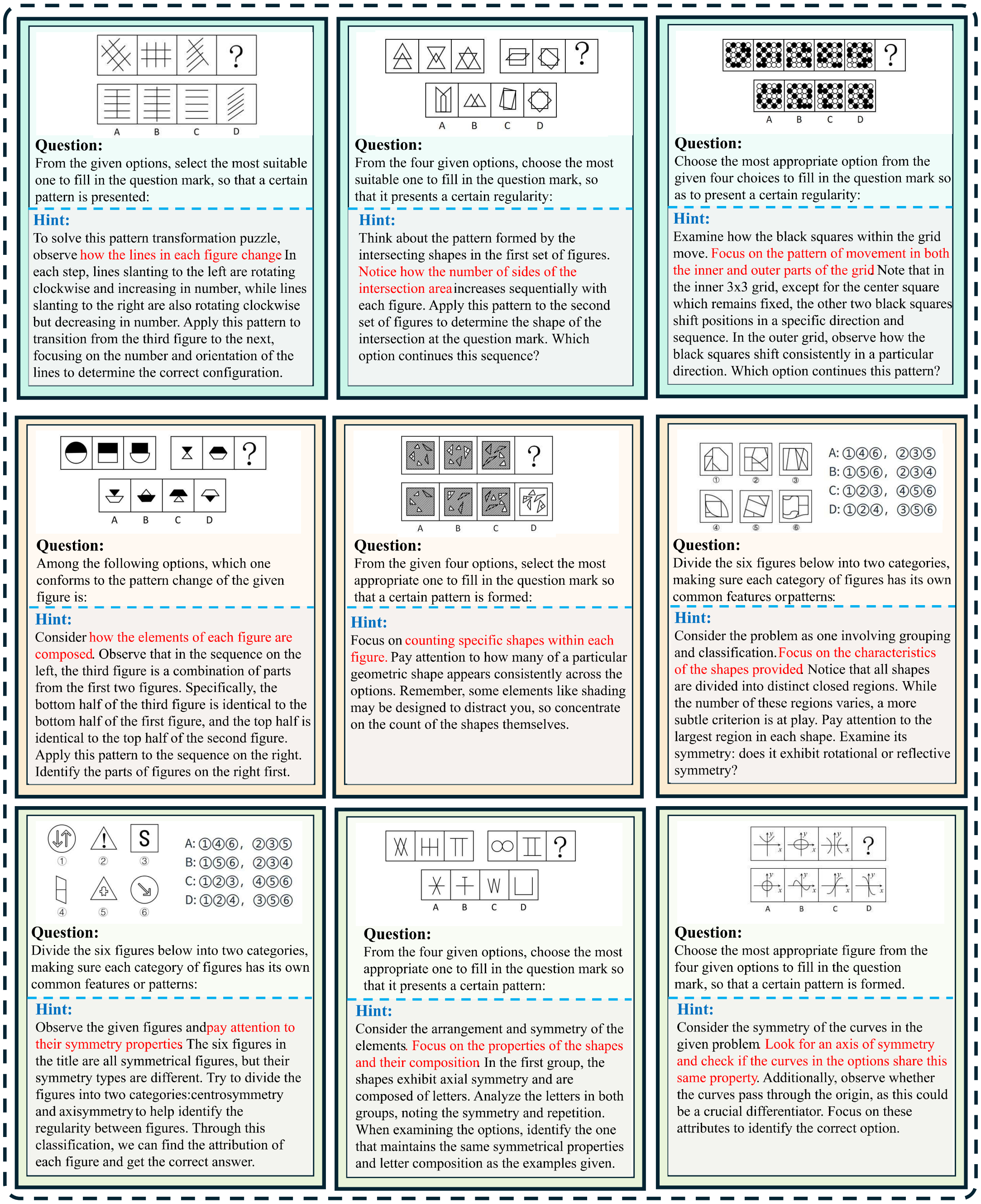
Gemini-2.0-pro-exp-02-05 ✓

Detailed step-by-step derivation:**\n1. **Analyze the sequence:** The given sequence consists of seven symbols followed by a question mark: `=`, `%`, `≈`, `÷`, `!`, `∴`, `:`, `?`.\n2. **Identify potential patterns:** Look for visual patterns, numerical patterns, or logical relationships between the symbols. One noticeable feature is the presence and number of dots in each symbol.\n3. **Count the dots:** Count the number of dots (or small circles) in each symbol:\n * `=` : 0 dots\n * `%` : 2 dots\n * `≈` : 0 dots\n * `÷` : 2 dots\n * `!` : 1 dot\n * `∴` : 3 dots\n * `:` : 2 dots\n * `?` : Unknown\n4. **Formulate the dot sequence:** The sequence of the number of dots is 0, 2, 0, 2, 1, 3, 2, ?.\n5. **Look for patterns within the dot sequence:** The sequence 0, 2, 0, 2, 1, 3, 2 doesn't immediately suggest a simple arithmetic or geometric progression. Let's examine alternating positions:\n * Odd positions (1st, 3rd, 5th, 7th): 0, 0, 1, 2. This sequence appears to be non-decreasing (0, 0, 1, 2).\n * Even positions (2nd, 4th, 6th, 8th): 2, 2, 3, ?. This sequence starts with two 2s, then a 3. A plausible continuation would be another 3, forming the pattern (2, 2, 3, 3).\n6. **Determine the required number of dots:** Based on the alternating pattern hypothesis, the 8th symbol (an even position) should follow the pattern (2, 2, 3, 3), meaning it should have 3 dots.\n7. **Evaluate the options:** Check the number of dots in each of the given options:\n * A. `#` : 0 dots\n * B. `∵` : 3 dots\n * C. `+` : 0 dots\n * D. `×` : 0 dots\n8. **Select the correct option:** Only option B (`∵`) has the required 3 dots.\n9. **Confirm the pattern:** If B is correct, the complete dot sequence is 0, 2, 0, 2, 1, 3, 2, 3. The patterns for odd (0, 0, 1, 2) and even (2, 2, 3, 3) positions hold.\n\nAnswer: The final answer is $\\boxed{B}$

o3-mini ✓

To solve the puzzle, let's analyze the sequence of symbols and identify the underlying pattern:\n\n1. **Sequence Analysis**:\n - **= (Equal Sign)**: Represents equality.\n - **% (Percent Sign)**: Represents a ratio or proportion.\n - **≈ (Approximately Equal Sign)**: Indicates approximation.\n - **÷ (Division Sign with Dots)**: Represents division.\n - **! (Exclamation Mark)**: Often used to denote factorial in mathematics or emphasis in language.\n - **∴ (Therefore Symbol)**: Indicates a conclusion.\n - **: (Colon)**: Used to introduce a list or explanation.\n\n2. **Identifying the Pattern**:\n - The sequence transitions from mathematical symbols to logical and punctuation symbols.\n - After the \"therefore\" symbol (∴), which signifies a conclusion, the colon (:) is used to introduce an explanation or reason.\n - Following this logical progression, the next symbol should logically represent \"because,\" which is denoted by the **\"∵\" (Because Symbol)**.\n\n3. **Evaluating the Options**:\n - **Option A (#)**: Hashtag or number sign, unrelated to the pattern.\n - **Option B (:∵)**: Combines the colon with the \"because\" symbol, fitting the logical continuation.\n - **Option C (+)**: Plus sign, unrelated to the current logical flow.\n - **Option D (×)**: Multiplication sign, also unrelated to the current pattern.\n\n4. **Conclusion**:\n - **Option B (:∵)** best fits the established pattern by introducing the \"because\" symbol, which logically follows the \"therefore\" symbol and colon.\n\n**Answer:** \\boxed{B}

Figure 18: **Solution examples generated by different models.** Reference solution and outputs generated by GPT-4o, Qwen2.5VL-72B, Gemini-2.0-pro-exp-02-05 and Doubao-1.5-Vision-Pro-32k. Additionally, the image caption and solution from LLMs (o3-mini) are also illustrated.



Question:
From the given options, select the most suitable one to fill in the question mark, so that a certain pattern is presented:

Hint:
To solve this pattern transformation puzzle, observe how the lines in each figure change In each step, lines slanting to the left are rotating clockwise and increasing in number, while lines slanting to the right are also rotating clockwise but decreasing in number. Apply this pattern to transition from the third figure to the next, focusing on the number and orientation of the lines to determine the correct configuration.

Question:
From the four given options, choose the most suitable one to fill in the question mark, so that it presents a certain regularity:

Hint:
Think about the pattern formed by the intersecting shapes in the first set of figures. Notice how the number of sides of the intersection areaincreases sequentially with each figure. Apply this pattern to the second set of figures to determine the shape of the intersection at the question mark. Which option continues this sequence?

Question:
Choose the most appropriate option from the given four choices to fill in the question mark so as to present a certain regularity:

Hint:
Examine how the black squares within the grid move. Focus on the pattern of movement in both the inner and outer parts of the grid Note that in the inner 3x3 grid, except for the center square which remains fixed, the other two black squares shift positions in a specific direction and sequence. In the outer grid, observe how the black squares shift consistently in a particular direction. Which option continues this pattern?

Question:
Among the following options, which one conforms to the pattern change of the given figure is:

Hint:
Consider how the elements of each figure are composed. Observe that in the sequence on the left, the third figure is a combination of parts from the first two figures. Specifically, the bottom half of the third figure is identical to the bottom half of the first figure, and the top half is identical to the top half of the second figure. Apply this pattern to the sequence on the right. Identify the parts of figures on the right first.

Question:
From the given four options, select the most appropriate one to fill in the question mark so that a certain pattern is formed:

Hint:
Focus on counting specific shapes within each figure. Pay attention to how many of a particular geometric shape appears consistently across the options. Remember, some elements like shading may be designed to distract you, so concentrate on the count of the shapes themselves.

A:①④⑥，②③⑤
B:①⑤⑥，②③④
C:①②③，④⑤⑥
D:①②④，③⑤⑥

Question:
Divide the six figures below into two categories, making sure each category of figures has its own common features orpatterns:

Hint:
Consider the problem as one involving grouping and classification. Focus on the characteristics of the shapes provided Notice that all shapes are divided into distinct closed regions. While the number of these regions varies, a more subtle criterion is at play. Pay attention to the largest region in each shape. Examine its symmetry: does it exhibit rotational or reflective symmetry?

A:①④⑥，②③⑤
B:①⑤⑥，②③④
C:①②③，④⑤⑥
D:①②④，③⑤⑥

Question:
Divide the six figures below into two categories, making sure each category of figures has its own common features or patterns:

Hint:
Observe the given figures andpay attention to their symmetry properties The six figures in the title are all symmetrical figures, but their symmetry types are different. Try to divide the figures into two categories:centrosymmetry and axisymmetryto help identify the regularity between figures. Through this classification, we can find the attribution of each figure and get the correct answer.

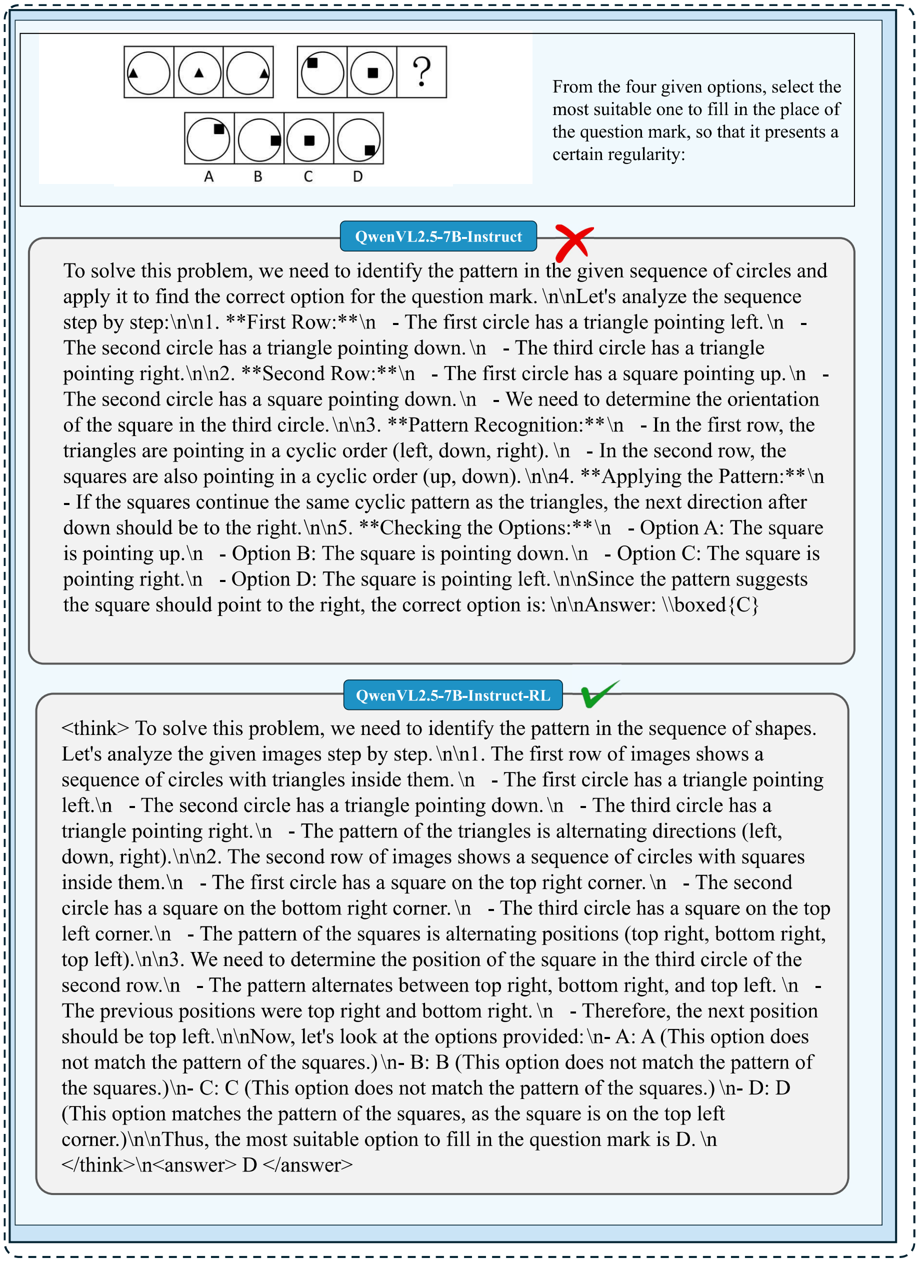
Question:
From the four given options, choose the most appropriate one to fill in the question mark so that it presents a certain pattern:

Hint:
Consider the arrangement and symmetry of the elements. Focus on the properties of the shapes and their composition In the first group, the shapes exhibit axial symmetry and are composed of letters. Analyze the letters in both groups, noting the symmetry and repetition. When examining the options, identify the one that maintains the same symmetrical properties and letter composition as the examples given.

Question:
Choose the most appropriate figure from the four given options to fill in the question mark, so that a certain pattern is formed.

Hint:
Consider the symmetry of the curves in the given problem. Look for an axis of symmetry and check if the curves in the options share this same property. Additionally, observe whether the curves pass through the origin, as this could be a crucial differentiator. Focus on these attributes to identify the correct option.

Figure 19: Examples of hint prompts. Hint prompts are provided to guide reasoning without revealing the final answer directly.



From the four given options, select the most suitable one to fill in the place of the question mark, so that it presents a certain regularity:

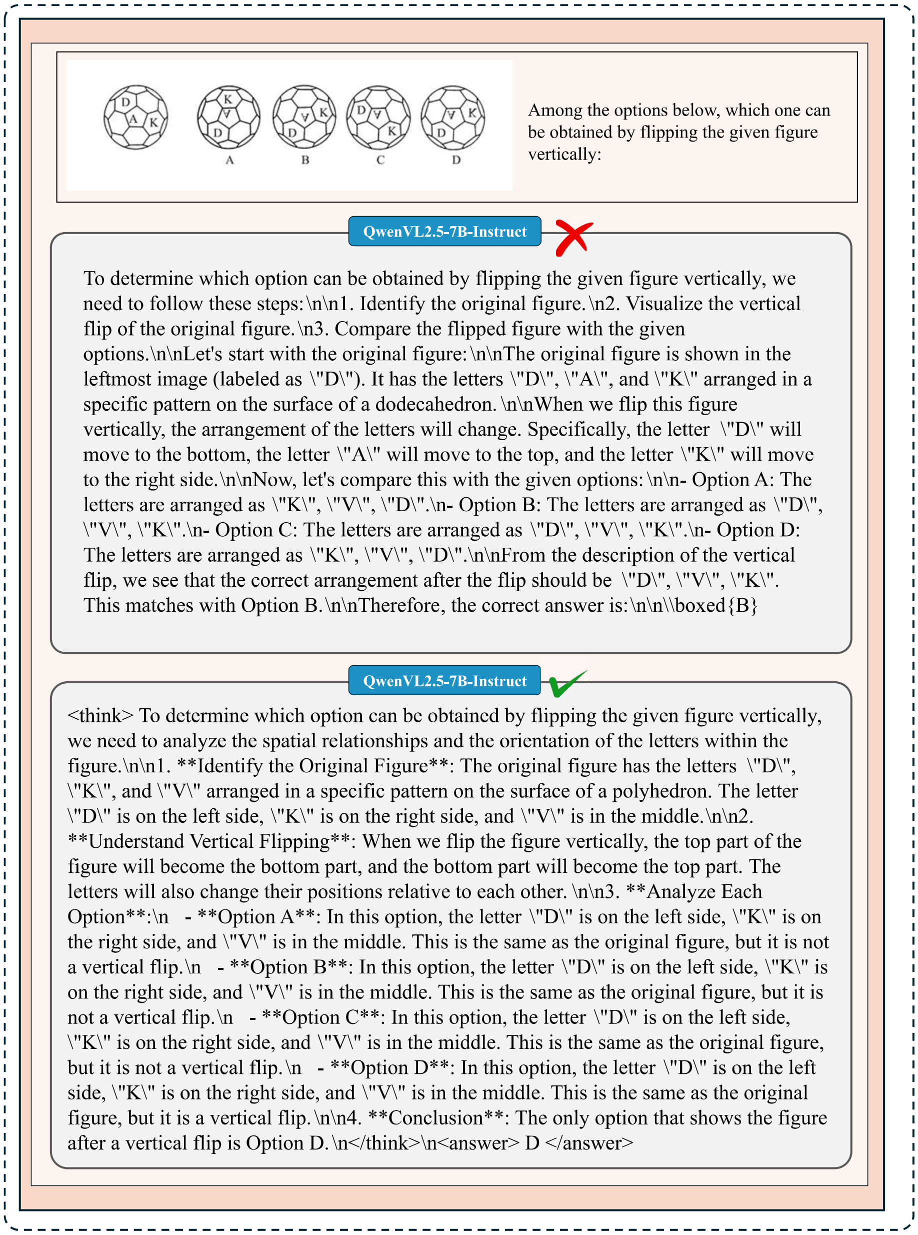
QwenVL2.5-7B-Instruct ✗

To solve this problem, we need to identify the pattern in the given sequence of circles and apply it to find the correct option for the question mark. \n\nLet's analyze the sequence step by step:\n\n1. **First Row:**\n - The first circle has a triangle pointing left. \n - The second circle has a triangle pointing down. \n - The third circle has a triangle pointing right.\n\n2. **Second Row:**\n - The first circle has a square pointing up. \n - The second circle has a square pointing down. \n - We need to determine the orientation of the square in the third circle. \n\n3. **Pattern Recognition:**\n - In the first row, the triangles are pointing in a cyclic order (left, down, right). \n - In the second row, the squares are also pointing in a cyclic order (up, down). \n\n4. **Applying the Pattern:**\n - If the squares continue the same cyclic pattern as the triangles, the next direction after down should be to the right.\n\n5. **Checking the Options:**\n - Option A: The square is pointing up.\n - Option B: The square is pointing down. \n - Option C: The square is pointing right.\n - Option D: The square is pointing left. \n\nSince the pattern suggests the square should point to the right, the correct option is: \n\nAnswer: \\boxed{C}

QwenVL2.5-7B-Instruct-RL ✓

<think> To solve this problem, we need to identify the pattern in the sequence of shapes. Let's analyze the given images step by step. \n\n1. The first row of images shows a sequence of circles with triangles inside them. \n - The first circle has a triangle pointing left.\n - The second circle has a triangle pointing down. \n - The third circle has a triangle pointing right.\n - The pattern of the triangles is alternating directions (left, down, right).\n\n2. The second row of images shows a sequence of circles with squares inside them.\n - The first circle has a square on the top right corner. \n - The second circle has a square on the bottom right corner. \n - The third circle has a square on the top left corner.\n - The pattern of the squares is alternating positions (top right, bottom right, top left).\n\n3. We need to determine the position of the square in the third circle of the second row.\n - The pattern alternates between top right, bottom right, and top left. \n - The previous positions were top right and bottom right. \n - Therefore, the next position should be top left.\n\nNow, let's look at the options provided:\n- A: A (This option does not match the pattern of the squares.)\n- B: B (This option does not match the pattern of the squares.)\n- C: C (This option does not match the pattern of the squares.) \n- D: D (This option matches the pattern of the squares, as the square is on the top left corner.)\n\nThus, the most suitable option to fill in the question mark is D. \n </think>\n<answer> D </answer>

Figure 20: Comparison of model outputs before and after RL training stage for Qwen2.5-VL-7B.



Among the options below, which one can be obtained by flipping the given figure vertically:

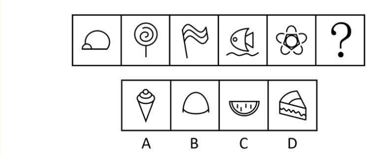
QwenVL2.5-7B-Instruct ✗

To determine which option can be obtained by flipping the given figure vertically, we need to follow these steps:\n\n1. Identify the original figure.\n2. Visualize the vertical flip of the original figure.\n3. Compare the flipped figure with the given options.\n\nLet's start with the original figure:\n\nThe original figure is shown in the leftmost image (labeled as \"D\"). It has the letters \"D\", \"A\", and \"K\" arranged in a specific pattern on the surface of a dodecahedron. \n\nWhen we flip this figure vertically, the arrangement of the letters will change. Specifically, the letter \"D\" will move to the bottom, the letter \"A\" will move to the top, and the letter \"K\" will move to the right side.\n\nNow, let's compare this with the given options:\n\n- Option A: The letters are arranged as \"K\", \"V\", \"D\".\n- Option B: The letters are arranged as \"D\", \"V\", \"K\".\n- Option C: The letters are arranged as \"D\", \"V\", \"K\".\n- Option D: The letters are arranged as \"K\", \"V\", \"D\".\n\nFrom the description of the vertical flip, we see that the correct arrangement after the flip should be \"D\", \"V\", \"K\". This matches with Option B.\n\nTherefore, the correct answer is:\n\n\\boxed{B}

QwenVL2.5-7B-Instruct ✓

<think> To determine which option can be obtained by flipping the given figure vertically, we need to analyze the spatial relationships and the orientation of the letters within the figure.\n\n1. **Identify the Original Figure**: The original figure has the letters \"D\", \"K\", and \"V\" arranged in a specific pattern on the surface of a polyhedron. The letter \"D\" is on the left side, \"K\" is on the right side, and \"V\" is in the middle.\n\n2. **Understand Vertical Flipping**: When we flip the figure vertically, the top part of the figure will become the bottom part, and the bottom part will become the top part. The letters will also change their positions relative to each other. \n\n3. **Analyze Each Option**:\n - **Option A**: In this option, the letter \"D\" is on the left side, \"K\" is on the right side, and \"V\" is in the middle. This is the same as the original figure, but it is not a vertical flip.\n - **Option B**: In this option, the letter \"D\" is on the left side, \"K\" is on the right side, and \"V\" is in the middle. This is the same as the original figure, but it is not a vertical flip.\n - **Option C**: In this option, the letter \"D\" is on the left side, \"K\" is on the right side, and \"V\" is in the middle. This is the same as the original figure, but it is not a vertical flip.\n - **Option D**: In this option, the letter \"D\" is on the left side, \"K\" is on the right side, and \"V\" is in the middle. This is the same as the original figure, but it is a vertical flip.\n\n4. **Conclusion**: The only option that shows the figure after a vertical flip is Option D.\n</think>\n<answer> D </answer>

Figure 21: Comparison of model outputs before and after RL training stage for Qwen2.5-VL-7B.



Choose the most appropriate option from the given four options to fill in the question mark, so that it presents a certain regularity:

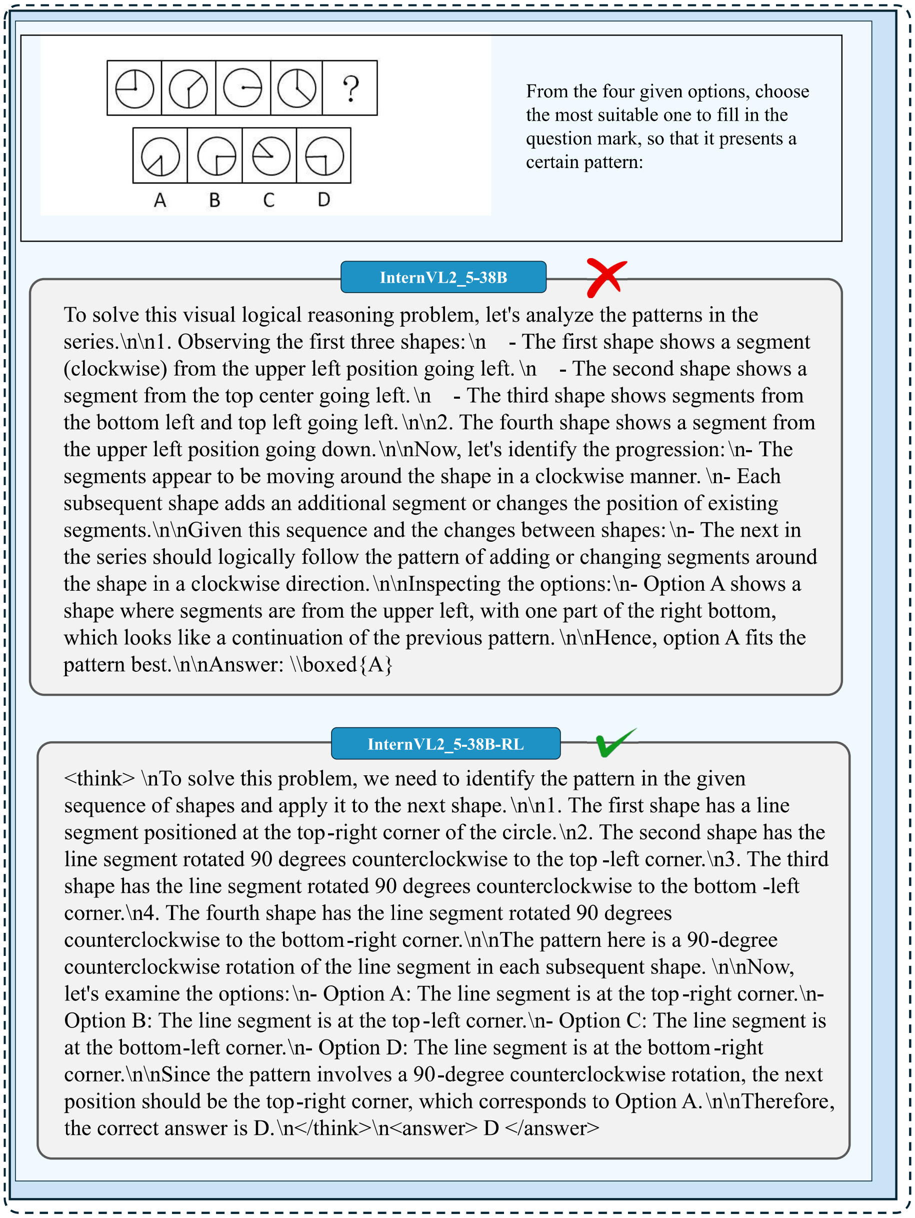
QwenVL2.5-7B-Instruct ✗

To solve this problem, we need to identify the pattern or rule that governs the sequence of images in the top row and then apply that same rule to find the correct image for the question mark.\n\nLet's analyze the sequence of images in the top row: \n\n1. The first image is a half-circle.\n2. The second image is a lollipop.\n3. The third image is a flag.\n4. The fourth image is a fish.\n5. The fifth image is an atom symbol.\n\nNow, let's look at the images in the bottom row:\n\nA: An ice cream cone.\nB: A half-circle.\nC: A slice of watermelon.\nD: A slice of cake.\n\nWe need to determine which of these images fits the pattern established by the top row. Let's consider the shapes and their transformations: \n\n- The first image (half-circle) is simple.\n- The second image (lollipop) has a spiral shape. \n- The third image (flag) has a wavy line.\n- The fourth image (fish) has a curved body.\n- The fifth image (atom symbol) has a complex structure with multiple parts. \n\nThe pattern seems to be the transformation of basic shapes into more complex forms. The half -circle becomes a lollipop, which has a spiral shape. The flag has a wavy line, and the fish has a curved body. The atom symbol is complex with multiple parts. \n\nLooking at the bottom row, the image that fits this pattern is the one that transforms a simple shape into a more complex form. The half-circle (image B) matches this description as it is the simplest shape in the top row and corresponds to the simplest shape in the bottom row. \n\nTherefore, the correct answer is:\n\nAnswer: \\boxed{B}

QwenVL2.5-7B-Instruct ✓

<think> To solve this problem, we need to identify the pattern in the sequence of images. Let's analyze the images step by step:\n\n1. The first image is a half-circle.\n2. The second image is a lollipop.\n3. The third image is a flag.\n4. The fourth image is a fish.\n5. The fifth image is a flower.\n6. The sixth image is a question mark, which we need to determine.\n\nNow, let's look at the options provided:\n\nA: An ice cream cone.\nB: A half-circle.\nC: A watermelon slice.\nD: A piece of cake.\n\nWe need to identify which of these options fits the pattern established by the previous images. Let's analyze the pattern: \n\n- The first image is a half-circle.\n- The second image is a lollipop, which has a circular shape but with a stick.\n- The third image is a flag, which has a distinct shape with a vertical and horizontal section.\n- The fourth image is a fish, which has a distinct shape with a tail and fins.\n- The fifth image is a flower, which has a distinct shape with petals and a center.\n\nThe pattern seems to be that each image has a distinct and recognizable shape. Now, let's look at the options:\n\n- Option A (ice cream cone) has a distinct shape with a cone and a scoop.\n- Option B (half-circle) is already seen in the first image. \n- Option C (watermelon slice) has a distinct shape with a triangular section. \n- Option D (piece of cake) has a distinct shape with a triangular section. \n\nSince the pattern is that each image has a distinct and recognizable shape, and the options A, C, and D all have distinct shapes, we need to choose the one that fits the pattern consistently. The most consistent and distinct shape among the options is the ice cream cone. \n\nTherefore, the most appropriate option to fill in the question mark is:\n <answer> A </answer>

Figure 22: Comparison of model outputs before and after RL training stage for Qwen2.5-VL-7B.



From the four given options, choose the most suitable one to fill in the question mark, so that it presents a certain pattern:

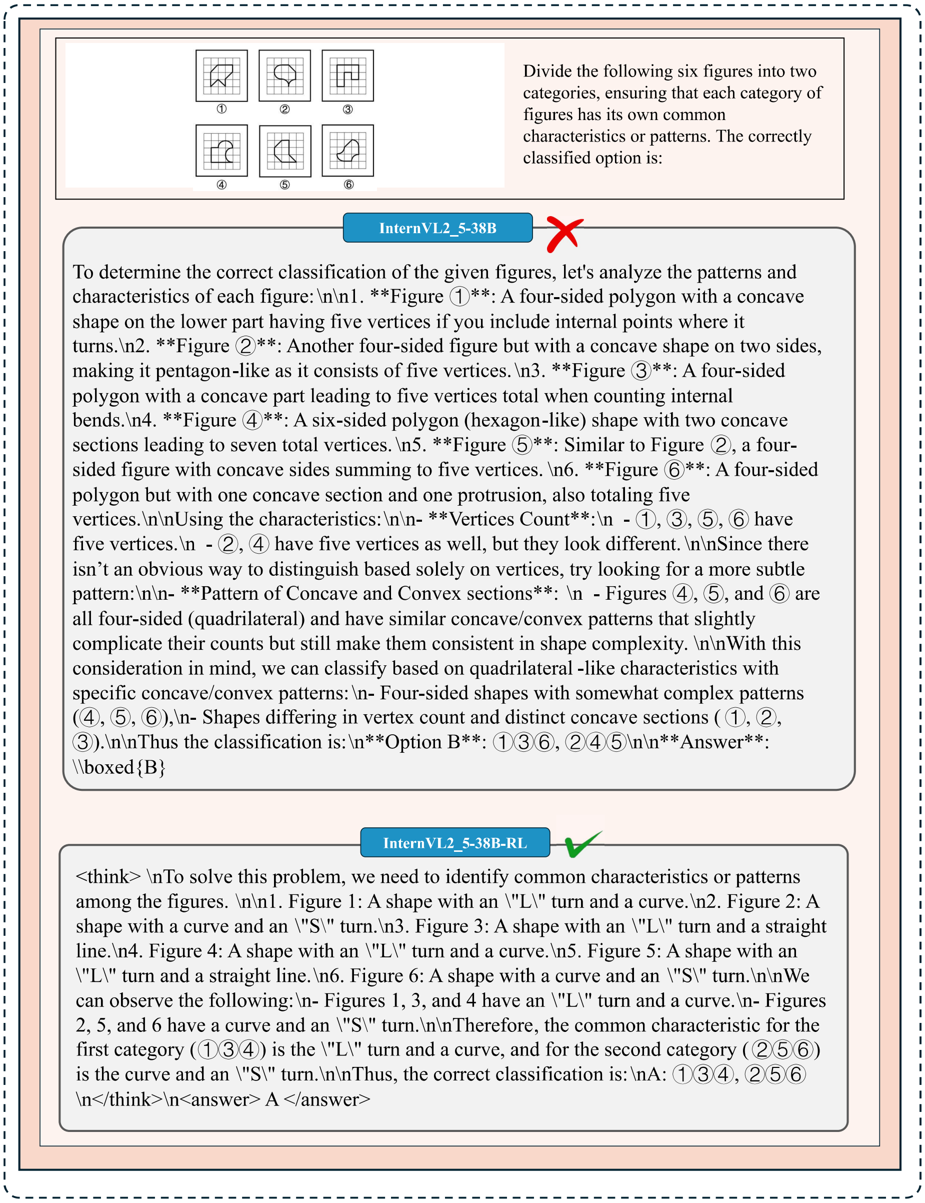
InternVL2_5-38B ✗

To solve this visual logical reasoning problem, let's analyze the patterns in the series.\n\n1. Observing the first three shapes:\n - The first shape shows a segment (clockwise) from the upper left position going left. \n - The second shape shows a segment from the top center going left.\n - The third shape shows segments from the bottom left and top left going left.\n\n2. The fourth shape shows a segment from the upper left position going down.\n\nNow, let's identify the progression:\n- The segments appear to be moving around the shape in a clockwise manner. \n- Each subsequent shape adds an additional segment or changes the position of existing segments.\n\nGiven this sequence and the changes between shapes:\n- The next in the series should logically follow the pattern of adding or changing segments around the shape in a clockwise direction. \n\nInspecting the options:\n- Option A shows a shape where segments are from the upper left, with one part of the right bottom, which looks like a continuation of the previous pattern. \n\nHence, option A fits the pattern best.\n\nAnswer: \\boxed{A}

InternVL2_5-38B-RL ✓

<think> \nTo solve this problem, we need to identify the pattern in the given sequence of shapes and apply it to the next shape. \n\n1. The first shape has a line segment positioned at the top-right corner of the circle.\n2. The second shape has the line segment rotated 90 degrees counterclockwise to the top -left corner.\n3. The third shape has the line segment rotated 90 degrees counterclockwise to the bottom -left corner.\n4. The fourth shape has the line segment rotated 90 degrees counterclockwise to the bottom-right corner.\n\nThe pattern here is a 90-degree counterclockwise rotation of the line segment in each subsequent shape. \n\nNow, let's examine the options:\n- Option A: The line segment is at the top-right corner.\n- Option B: The line segment is at the top-left corner.\n- Option C: The line segment is at the bottom-left corner.\n- Option D: The line segment is at the bottom-right corner.\n\nSince the pattern involves a 90-degree counterclockwise rotation, the next position should be the top-right corner, which corresponds to Option A. \n\nTherefore, the correct answer is D.\n</think>\n<answer> D </answer>

Figure 23: Comparison of model outputs before and after RL training stage for InternVL-2.5-38B.



Divide the following six figures into two categories, ensuring that each category of figures has its own common characteristics or patterns. The correctly classified option is:

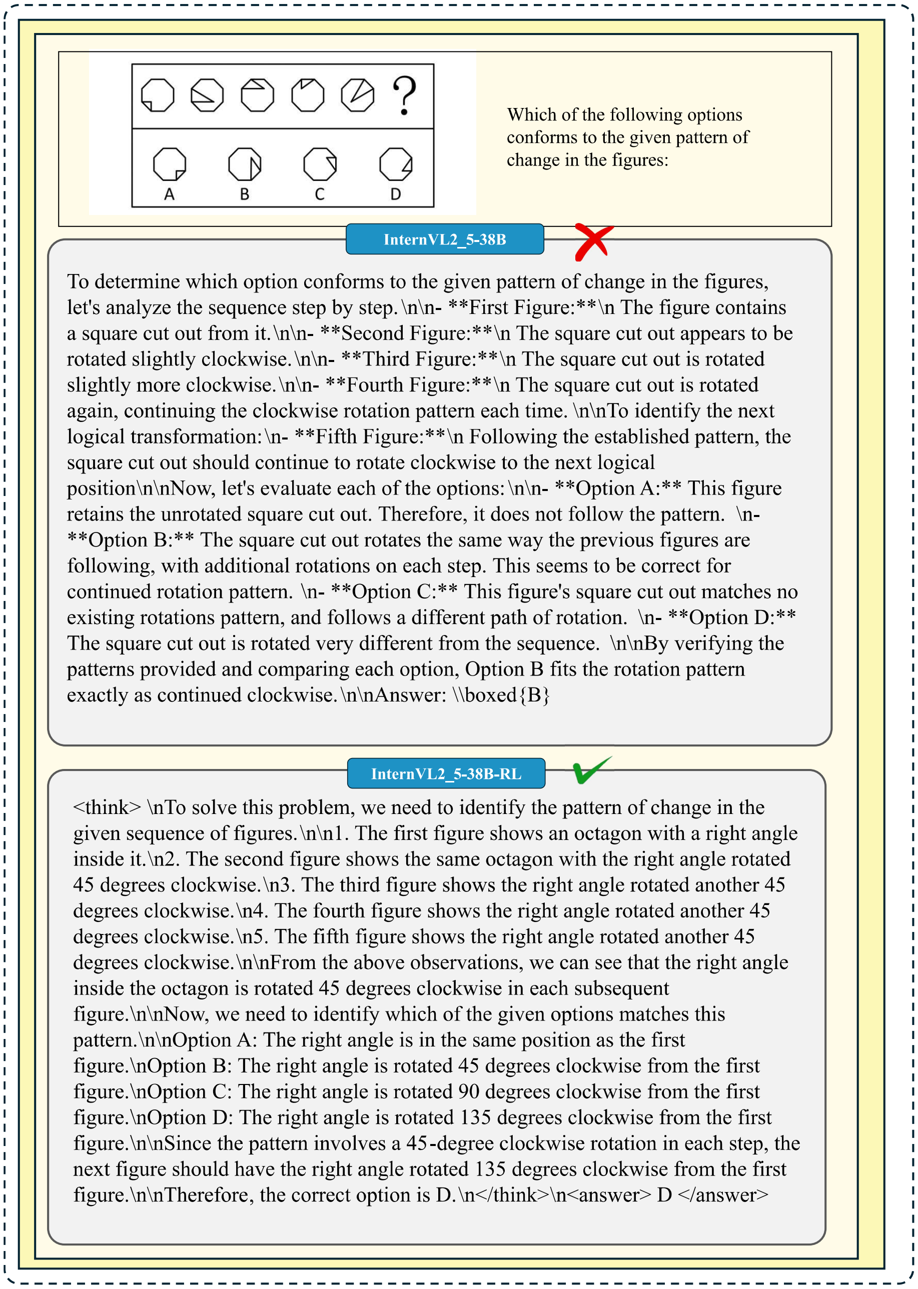
InternVL2_5-38B ✗

To determine the correct classification of the given figures, let's analyze the patterns and characteristics of each figure:\n\n1. **Figure ①**: A four-sided polygon with a concave shape on the lower part having five vertices if you include internal points where it turns.\n2. **Figure ②**: Another four-sided figure but with a concave shape on two sides, making it pentagon-like as it consists of five vertices.\n3. **Figure ③**: A four-sided polygon with a concave part leading to five vertices total when counting internal bends.\n4. **Figure ④**: A six-sided polygon (hexagon-like) shape with two concave sections leading to seven total vertices.\n5. **Figure ⑤**: Similar to Figure ②, a four-sided figure with concave sides summing to five vertices. \n6. **Figure ⑥**: A four-sided polygon but with one concave section and one protrusion, also totaling five vertices.\n\nUsing the characteristics:\n\n- **Vertices Count**:\n - ①, ③, ⑤, ⑥ have five vertices.\n - ②, ④ have five vertices as well, but they look different. \n\nSince there isn't an obvious way to distinguish based solely on vertices, try looking for a more subtle pattern:\n\n- **Pattern of Concave and Convex sections**: \n - Figures ④, ⑤, and ⑥ are all four-sided (quadrilateral) and have similar concave/convex patterns that slightly complicate their counts but still make them consistent in shape complexity. \n\nWith this consideration in mind, we can classify based on quadrilateral -like characteristics with specific concave/convex patterns:\n- Four-sided shapes with somewhat complex patterns (④, ⑤, ⑥),\n- Shapes differing in vertex count and distinct concave sections (①, ②, ③).\n\nThus the classification is:\n**Option B**: ①③⑥, ②④⑤\n\n**Answer**: \\boxed{B}

InternVL2_5-38B-RL ✓

<think> \nTo solve this problem, we need to identify common characteristics or patterns among the figures. \n\n1. Figure 1: A shape with an \"L\" turn and a curve.\n2. Figure 2: A shape with a curve and an \"S\" turn.\n3. Figure 3: A shape with an \"L\" turn and a straight line.\n4. Figure 4: A shape with an \"L\" turn and a curve.\n5. Figure 5: A shape with an \"L\" turn and a straight line.\n6. Figure 6: A shape with a curve and an \"S\" turn.\n\nWe can observe the following:\n- Figures 1, 3, and 4 have an \"L\" turn and a curve.\n- Figures 2, 5, and 6 have a curve and an \"S\" turn.\n\nTherefore, the common characteristic for the first category (①③④) is the \"L\" turn and a curve, and for the second category (②⑤⑥) is the curve and an \"S\" turn.\n\nThus, the correct classification is:\nA: ①③④, ②⑤⑥ \n</think>\n<answer> A </answer>

Figure 24: Comparison of model outputs before and after RL training stage for InternVL-2.5-38B.



Which of the following options conforms to the given pattern of change in the figures:

InternVL2_5-38B ✗

To determine which option conforms to the given pattern of change in the figures, let's analyze the sequence step by step. \n\n- **First Figure:**\n The figure contains a square cut out from it. \n\n- **Second Figure:**\n The square cut out appears to be rotated slightly clockwise. \n\n- **Third Figure:**\n The square cut out is rotated slightly more clockwise. \n\n- **Fourth Figure:**\n The square cut out is rotated again, continuing the clockwise rotation pattern each time. \n\nTo identify the next logical transformation: \n- **Fifth Figure:**\n Following the established pattern, the square cut out should continue to rotate clockwise to the next logical position\n\nNow, let's evaluate each of the options: \n\n- **Option A:** This figure retains the unrotated square cut out. Therefore, it does not follow the pattern. \n- **Option B:** The square cut out rotates the same way the previous figures are following, with additional rotations on each step. This seems to be correct for continued rotation pattern. \n- **Option C:** This figure's square cut out matches no existing rotations pattern, and follows a different path of rotation. \n- **Option D:** The square cut out is rotated very different from the sequence. \n\nBy verifying the patterns provided and comparing each option, Option B fits the rotation pattern exactly as continued clockwise. \n\nAnswer: \\boxed{B}

InternVL2_5-38B-RL ✓

<think> \nTo solve this problem, we need to identify the pattern of change in the given sequence of figures. \n\n1. The first figure shows an octagon with a right angle inside it. \n2. The second figure shows the same octagon with the right angle rotated 45 degrees clockwise. \n3. The third figure shows the right angle rotated another 45 degrees clockwise. \n4. The fourth figure shows the right angle rotated another 45 degrees clockwise. \n5. The fifth figure shows the right angle rotated another 45 degrees clockwise. \n\nFrom the above observations, we can see that the right angle inside the octagon is rotated 45 degrees clockwise in each subsequent figure. \n\nNow, we need to identify which of the given options matches this pattern. \n\nOption A: The right angle is in the same position as the first figure. \nOption B: The right angle is rotated 45 degrees clockwise from the first figure. \nOption C: The right angle is rotated 90 degrees clockwise from the first figure. \nOption D: The right angle is rotated 135 degrees clockwise from the first figure. \n\nSince the pattern involves a 45-degree clockwise rotation in each step, the next figure should have the right angle rotated 135 degrees clockwise from the first figure. \n\nTherefore, the correct option is D. \n</think>\n<answer> D </answer>

Figure 25: Comparison of model outputs before and after RL training stage for InternVL-2.5-38B.