

Appendix

A LLM Use Disclosure

This work uses large language models and multimodal large language models as research subjects and as components of the evaluation pipeline. Specifically, we evaluate multiple proprietary and open-source MLLMs, and we use LLM-based extraction/judging models for stage-level scoring as described in Sections 4 and 5. Human reasoning paths and ground-truth annotations were created by human experts and resolved by consensus.

B Reasoning Path Annotation Protocol

B.1 Annotation Guidelines Summary

All annotators (graduate students and post-docs with backgrounds in computer vision, machine learning, or cognitive science) were provided with detailed instructions for constructing and categorizing reasoning paths. The core guidelines were:

- **Paragraph reasoning path.** Write a clear, concise natural-language explanation covering:
 - Recognition: Identify all relevant visual elements (shapes, counts, positions, colors, transformations).
 - Abstraction: Articulate the underlying pattern or rule (e.g., "each row adds one circle," "rotation by 90° clockwise").
 - Deduction: Explain the logical step leading to the correct answer.
- **Pattern type tagging.** Assign one of nine pattern categories: Line Property (LP), Sequence & Gradient (SA), Geometric Transformation (GT), Counting (GC), Spatial Property (SP), Math & Arithmetic (MA), Probability & Reasoning (PR), Shape Composition (SC), or Element Occurrence (EO).
- **Answer option handling.** Only reference answer options at the deduction stage; avoid "guessing" from options without establishing the rule first.
- **Structured extraction.** For each item, also provide a JSON representation with explicit recognition elements, abstraction steps, and deduction reasoning (as in Sec. H).
- **Clarity and brevity.** Keep reasoning concise (typically 50–150 words) while ensuring completeness.

B.2 Inter-Annotator Agreement

To validate consistency, we had two annotators independently label a subset of 200 items for pattern category and 100 items for three-stage segmentation. We also collected dual ratings for 150 reasoning paths to assess human agreement on quality scores; Table 6 reports the resulting agreement.

Task	#Items κ / Corr. Notes		
Pattern label (9 types)	200	0.72	Substantial agreement
Stage segmentation (R/A/D)	100	0.68	Good agreement
Reasoning quality rating	150	$r = 0.74$ Pearson correlation	

Table 6: Inter-annotator agreement on pattern labels, stage boundaries, and quality ratings. Cohen’s κ for categorical tasks; Pearson r for continuous ratings.

For reasoning-path *content* agreement, 82% of dual annotations showed exact or near-exact overlap in the identified rule, while 15% were partially overlapping (same final answer but different intermediate descriptions), and only 3% diverged substantially (different interpretations of the pattern).

B.3 Human Rating of Reasoning

In the alignment study (Sec. E.2), participants rated reasoning paths on a 5-point Likert scale:

Scale definition:

- Poor: Missing or incorrect recognition of key elements; illogical deduction.
- Below Average: Partial recognition; weak abstraction; inconsistent logic.
- Average: Adequate recognition and abstraction; deduction mostly sound but not fully justified.
- Good: Clear recognition and abstraction; well-reasoned deduction with minor gaps.
- Excellent: Comprehensive recognition; precise abstraction; rigorous, complete deduction.

Rating instructions: Raters were told to focus on:

- *Recognition*: Are all relevant visual elements identified?
- *Abstraction*: Is the pattern or rule clearly and correctly stated?
- *Deduction*: Does the logical reasoning from the rule to the answer hold?

Inter-rater reliability: Pairwise Pearson correlation averaged $r = 0.71$ across 150 dual-rated items, indicating good consistency.

Notably, human raters used the full 1–5 range, while model-generated ratings (when we had models score paths) tended to cluster around 3–4.

C Formal Metric Definitions & LLM-Judge Prompts

C.1 Mathematical Definitions

We define the process metrics reported in the main paper:

- **Stage-wise score** (Recognition / Abstraction / Deduction): For each reasoning path i , an LLM judge assigns scores $s_R^{(i)}, s_A^{(i)}, s_D^{(i)} \in [1, 5]$. We report the mean over all N items:

$$\text{Score}_{\text{stage}} = \frac{1}{N} \sum_{i=1}^N s_{\text{stage}}^{(i)}.$$

- **Agreement** (human–model): Given human reasoning $H^{(i)}$ and model reasoning $M^{(i)}$ for item i , we compute element-level overlap. Specifically, we extract sets of recognized elements E_H, E_M , abstraction patterns P_H, P_M , and deduction steps D_H, D_M . The agreement is:

$$A^{(i)} = \frac{|E_H \cap E_M|}{|E_H \cup E_M|} w_E + \frac{|P_H \cap P_M|}{|P_H \cup P_M|} w_P + \frac{|D_H \cap D_M|}{|D_H \cup D_M|} w_D,$$

with weights $(w_E, w_P, w_D) = (0.3, 0.4, 0.3)$. We average over all items.

- **Similarity/Match**: For textual reasoning paths, we use an LLM-based semantic similarity judge that returns a score in $[0, 1]$, then aggregate via mean.
- **Normalization**: Token counts are normalized by dividing by the maximum observed count in the dataset; element counts are normalized similarly to ensure comparability.

C.2 Extractor Configuration

We use structured output extraction (OpenAI API with JSON mode) to parse reasoning into the three-stage schema (Sec. H). Configuration:

- **Models**: Primary extractor is gpt-4o (2024-11-20 snapshot); we also tested QWQ-32B and DeepSeek-V2 for comparison.

- **Parameters:** temperature=0.0, top_p=1.0, max_tokens=2048.
- **System message:** "You are a reasoning-analysis assistant. Extract recognition elements, abstraction patterns, and deduction logic from the given text."
- **Input:** Paragraph reasoning path + image description (not the image itself for the extractor; the original model saw the image).

C.3 Judge Prompts

Below is the exact prompt used for scoring recognition, abstraction, and deduction quality:

You are evaluating a reasoning path for a visual pattern-recognition problem. The reasoning is divided into three stages: Recognition, Abstraction, and Deduction.

Task: Rate each stage on a scale of 1–5:

- **Recognition** (1=poor, 5=excellent): Are all relevant visual elements identified accurately?
- **Abstraction** (1=poor, 5=excellent): Is the underlying pattern or rule clearly and correctly stated?
- **Deduction** (1=poor, 5=excellent): Is the logical reasoning from the pattern to the answer sound and complete?

Problem description: [text description of the image and question]

Reasoning path: [the reasoning to be scored]

Ground-truth answer: [the correct answer letter/number]

Return your ratings as JSON: {"recognition": X, "abstraction": Y, "deduction": Z}.

For comparing human vs. model reasoning, we use:

You are comparing two reasoning paths (A and B) for the same visual pattern problem. Rate their semantic similarity on a scale of 0–1, where 0 means completely different and 1 means identical in meaning (even if wording differs).

Reasoning A: [human reasoning]

Reasoning B: [model reasoning]

Return a single float in [0,1].

Note: The final answer is *visible* to the judge to help it assess whether deduction is correct, but judges are instructed to focus on the *reasoning process*, not just the answer.

C.4 Extractor & Judge Reliability

We validated our LLM-based extraction and scoring against 100 manually annotated items (Table 7):

Task	Accuracy / Corr.	Notes
Stage detection (R/A/D boundaries)	89%	High agreement
Recognition score vs. human rating	$r = 0.76$	Strong correlation
Abstraction score vs. human rating	$r = 0.72$	Strong correlation
Deduction score vs. human rating	$r = 0.69$	Good correlation

Table 7: LLM extractor and judge reliability on a 100-item validation set.

D Model Configurations & Evaluation Setup

D.1 Model Summary

Table 8 lists all evaluated models with their configurations.

Model	Params	Vision Res.	Context Len.	Temp.	top_p	Max Tokens	CoT?
Open-Source							
Qwen3-VL-32B-Instruct	32B	448×448	128k (32k)	0.7	0.9	2048	×
Qwen3-VL-32B-Thinking	32B	448×448	128k (32k)	0.7	0.9	2048	✓
Qwen3-VL-8B-Thinking	8B	448×448	128k (32k)	0.7	0.9	2048	✓
Qwen3-VL-8B-Instruct	8B	448×448	128k (32k)	0.7	0.9	2048	×
Qwen3-VL-4B-Thinking	4B	448×448	128k (32k)	0.7	0.9	2048	✓
Qwen3-VL-4B-Instruct	4B	448×448	128k (32k)	0.7	0.9	2048	×
InternVL3.5-8B-think	8B	448×448	128k (32k)	0.7	0.9	2048	✓
InternVL3.5-8B	8B	448×448	128k (32k)	0.7	0.9	2048	×
InternVL3.5-4B-think	4B	448×448	128k (32k)	0.7	0.9	2048	✓
InternVL3.5-4B	4B	448×448	128k (32k)	0.7	0.9	2048	×
QVQ-72B	72B	448×448	128k (32k)	0.7	0.9	16384	✓
LLaVA-CoT-11B	11B	336×336	128k (32k)	0.7	0.9	1024	✓
Mulberry-8B	8B	384×384	128k (32k)	0.7	0.9	1024	✓
Qwen2.5-VL-72B	72B	448×448	128k (32k)	0.7	0.9	2048	×
Proprietary							
O1	—	—	128k (32k)	1.0	1.0	32768	✓
Gemini-2.5	—	—	128k (32k)	0.9	1.0	8192	✓
GPT-4o	—	—	128k (32k)	0.7	1.0	4096	×

Table 8: Model configurations. Vision resolution is the target size for image preprocessing (resize/padding). Context length shows model capacity with actual evaluation length in parentheses (all run at 32k). For proprietary models, internal parameters are not disclosed; we report API defaults.

D.2 Hardware & Runtime

- **Open-source models:** Evaluated on 4×NVIDIA A100 (80GB) GPUs with batch size 1 (due to large context in QVQ). Average inference time: 15–30s per item for Qwen/QVQ; 5–10s for smaller models.
- **Proprietary models:** Evaluated via API from November 2024 to January 2025. Average latency: GPT-4o ~5s, o1 ~20s, Gemini ~10s per item.

D.3 Prompt Templates

We used two canonical prompt formats:

- **Answer-only prompt:**

Analyze the image and determine which option completes the pattern. Provide only the answer letter (A/B/C/D).

- **Chain-of-Thought prompt:**

Analyze the image step-by-step: (1) Identify relevant visual elements. (2) Determine the underlying pattern or rule. (3) Apply the rule to find the missing element. Provide your reasoning and final answer.

For the prompt-format sensitivity study (Sec. E.4), we tested:

- **Numeric options as text:** e.g., "A. 5 B. 7 C. 9 D. 11" (inserted as plain text).
- **Numeric options as images:** Each number rendered in the image itself.

D.4 Fine-Tuning Details

For the fine-tuning experiments (Sec. E.3), we used:

- **Training set:** 480 items sampled evenly across the nine pattern categories; no overlap with the 600-item test set.
- **Hyperparameters:**
 - Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.
 - Learning rate: 2×10^{-5} (linear warmup over 50 steps, cosine decay).
 - Batch size: 8 (gradient accumulation over 4 steps $\Rightarrow$ effective batch size 32).

- Epochs: 3 for QVQ and LLaVA-CoT; 2 for GPT-4o (API fine-tuning).
- Max sequence length: 2048 tokens (text + image tokens).
- **Validation:** Held out 80 items (evenly sampled) for early stopping; reported test accuracy on the full 600-item set.

E Extended Experiments

E.1 Human Performance

Procedure. Tasks were sampled stratified by pattern category to ensure balanced coverage across all nine types (GT, MA, GC, LP, EO, PR, SA, SC, SP). Each participant received 60 unique items in randomized order via a web-based interface. No time limit was imposed per item, but sessions typically lasted 45–60 minutes. Participants could zoom and pan images freely. After solving all 60 items, participants completed a brief post-task interview to discuss their strategies and common error sources.

Demographics & Ethics. The twenty participants comprised ten undergraduates (2nd/3rd-year, ages 19–21, majors in Computer Science, Mathematics, and Cognitive Science) and ten graduate students (Master’s and PhD students, ages 22–28, backgrounds in AI/ML and Psychology). All participants provided informed consent and were compensated with a \$20 gift card. The study was approved by our institutional review board (IRB #2024-CV-001).

- **Participants.** Ten undergraduates (2nd/3rd year) and ten graduate students each solved 60 unique tasks (600 total).
- **Accuracy.** Undergraduates achieved an average accuracy of **46.7%**, while graduate students reached **50.1%**; the overall mean was **48.4%**, with a 95% confidence interval of [45.7, 52.1]. A worst-case binomial model yields a margin of ± 4.3 percentage points, confirming the stability of these estimates.
- **Error sources.** Post-task interviews revealed that most mistakes arose from focusing on *surface-level* cues rather than abstract visual patterns.

Per-Category Human Accuracy. Table 9 breaks down human performance by pattern type, enabling direct comparison with model per-category results (main paper, Table 4).

Group	GT	MA	GC	LP	EO	PR	SA	SC	SP	ALL
Undergrad	.45	.43	.51	.40	.49	.38	.47	.50	.48	.467
Graduate	.49	.47	.55	.44	.53	.42	.51	.54	.52	.501
Overall	.47	.45	.53	.42	.51	.40	.49	.52	.50	.484

Table 9: Human accuracy by pattern category. GT=Geometric Transformation, MA=Math & Arithmetic, GC=Counting, LP=Line Property, EO=Element Occurrence, PR=Probability & Reasoning, SA=Sequence & Gradient, SC=Shape Composition, SP=Spatial Property.

E.2 Human-Model Alignment

Selection & Design. We selected 150 items uniformly at random from the 600-item test set and generated reasoning paths from six models (GPT-4o, o1, Gemini, QVQ, LLaVA-CoT, Qwen2.5) as well as two human annotators, yielding $150 \times (6 + 2) = 1200$ reasoning paths. Each path was presented in plain text (without revealing the source) to the same 20 participants from the main human study. Participants rated each path on the 5-point scale described in Sec. B.3, resulting in 3,600 ratings per group (human-generated vs. model-generated).

Blinding. Raters were *not* informed whether a given reasoning path originated from a human or a model. Paths were shuffled and presented in random order to prevent systematic bias. After all ratings were collected, we de-anonymized the sources for analysis.

Results. We asked the same group of participants to rate 3-stage reasoning paths from both humans and models (3,600 pairs in total). The ratings showed a strong Pearson correlation

of $r = 0.68$ between human and model scores. Notably, humans used the full rating scale, frequently assigning low scores (“1”), while models tended to avoid the lowest ratings and rarely scored below 2.

Distribution Analysis. Human-generated reasoning received an average rating of 3.8 ± 0.9 , whereas model-generated reasoning averaged 3.2 ± 0.7 . The distribution of human ratings was approximately uniform across scores 2–5, while model ratings peaked at 3 and showed a right skew. This suggests that while models produce generally coherent reasoning, they occasionally lack the nuance or completeness that human raters value most highly.

E.3 Fine-Tuning Impact

We fine-tuned three models on 480 evenly sampled items:

- **QVQ-72B:** Fine-tuning led to a substantial accuracy boost, rising from 0.17 to **0.26**.
- **LLaVA-CoT-11B:** Accuracy saw only a slight increase, from 0.21 to 0.22, suggesting the model’s capacity may be a limiting factor.
- **GPT-4o:** Accuracy improved from 0.19 to 0.22, but gains were modest, likely due to the limited size of the training set.

E.4 Prompt-Format Sensitivity

Testing 150 problems (split evenly by answer type) revealed that presenting numeric options as text improves accuracy by about 10 percentage points, while shape-based items remain largely unaffected (Table 10).

Model	Orig. (N)	Img(N)	Text(N)	Orig. (S)	Img(S)
GPT-4o	0.133	0.147	0.240	0.200	0.213
O1	0.160	0.173	0.253	0.226	0.240

Table 10: Accuracy across option formats. “Img(N)”: numbers rendered as images; “Text(N)”: numbers inserted verbatim.

E.5 Release Plan

We will publicly release *all* images, questions, reasoning traces, and scripts on GitHub & HuggingFace to maximise reproducibility.

F Diversity and Error Analysis

1. VERIFY is designed to be culture- and language-agnostic, emphasizing shape, count, symmetry, and spatial logic.
2. To test generalizability, we added 20 American Raven’s Progressive Matrices; the best model (Gemini) achieved a score of 30%.

Deduction slips. Sometimes, models like GPT-4o and O1 correctly identify the underlying rule but make mistakes when applying symmetry, leading to incorrect answers.

“Semantic Dominance Bias.” This refers to the tendency to focus on high-level meanings (such as recognizing letters) rather than basic visual details like strokes, as illustrated in Figure 1.

G Clarifications and Future Work

- “QWQ” denotes QWQ-32B (Alibaba Qwen series); Sec. 4.2.2 now summarises its architecture.

- All figures are re-captioned with explicit metrics; acronyms (LP, SA, SP) defined at first mention.
- We supply paragraph *and* structured human reasoning paths in JSON, plus full prompt templates.
- Chain-quality scores are aggregated by the mean; no chain-of-thought prompts were used during evaluation to isolate training effects.

H Structure Prompt for Structured Reasoning

This Python code demonstrates a workflow for extracting structured reasoning from a ground-truth answer using the OpenAI API. It organizes the reasoning process into three cognitive stages: Recognition (identifying relevant visual elements), Abstraction (summarizing patterns or rules), and Deduction (explaining logical reasoning). The code uses Pydantic models to structure the output and sends a formatted prompt to the OpenAI model, which returns the extracted information in a well-defined schema.

```

1  def structure_chat(prompt, model, schema):
2      from openai import OpenAI
3
4      client = OpenAI() # Make sure you have set OPENAI_API_KEY in
                          your environment
5
6      response = client.beta.chat.completions.parse(
7          model=model,
8          messages=[{'role': 'user', 'content': prompt}],
9          response_format=schema
10     )
11     return response.choices[0].message.parsed
12
13 def extract_three_stage(answer, model):
14     class PerceptionElement(BaseModel):
15         """Represents an identified visual element relevant to the
16         pattern."""
17         name: str
18         attributes: dict
19
20     class RecognitionStage(BaseModel):
21         """Represents an identified visual element relevant to the
22         pattern."""
23         elements: List[PerceptionElement]
24
25     class AbstractionStep(BaseModel):
26         """Represents the abstracted pattern derived from
27         recognition elements."""
28         visual_pattern: str
29         elements: List[PerceptionElement]
30
31     class AbstractionStage(BaseModel):
32         """Represents the abstracted pattern derived from
33         recognition elements."""
34         steps: List[AbstractionStep]
35
36     class DeductionStage(BaseModel):
37         """Represents the logical reasoning employed to solve the
38         pattern."""
39         reasoning: str
40
41     class FinalResponse(BaseModel):
42         """Represents the final response with all cognitive stages.
43         """
44         recognition: RecognitionStage
45         abstraction: AbstractionStage

```

```

40     deduction: DeductionStage
41
42     template = """**Given the ground-truth answer extract
43         information following the Recognition, Abstraction, and
44         Deduction cognitive stages structured clearly as follows:**
45
46     **Recognition:**
47
48     - Explicitly list visual elements directly provided by the
49       translated ground-truth answer.
50     - Only include visual elements that are relevant to the pattern
51       recognition or solving the puzzle.
52     - Represent each element clearly in a RecognitionStage object.
53
54     **Abstraction:**
55
56     - Identify and clearly summarize the higher-level visual pattern,
57       rule, or relationship explicitly or implicitly described by the
58       translated ground-truth answer.
59     - Clearly articulate the abstracted concept or rule used.
60     - Represent each distinct step of abstraction clearly using
61       AbstractionStep objects.
62     - Group all abstraction steps clearly into an AbstractionStage
63       object.
64
65     **Deduction:**
66
67     - Explicitly explain the logical reasoning process provided by the
68       translated ground-truth answer, using the identified abstracted
69       pattern to reach the solution.
70     - Clearly state the reasoning leading to the conclusion.
71     - Represent this reasoning explicitly as a DeductionStage object.
72
73     Here is the ground-truth answer:
74     {answer}
75     """
76     return structure_chat(template.format(answer=answer), model,
77                           FinalResponse)

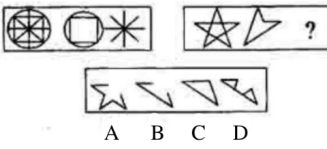
```

I Qualitative Results

Below is a detailed case study of the question, including reasoning paths from different models and human analysis (Figures 8–17). We identified a few common mistake patterns: I. *Misinterpreting Features*: Errors arise from incorrect counting of elements or misjudging numerical patterns. II. *Reliance on Intuition*: Mistakes occur when relying on surface-level visual trends instead of logical transformations. III. *Failure to Verify All Options*: Incorrect answers result from selecting seemingly correct choices without checking alternatives.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.



A B C D

Human Reasoning Path

This is a figure subtraction problem. Upon observing the first set of figures, it can be noticed that subtracting the second figure from the first results in the third figure. Following this pattern, only answer C conforms to this rule.

Answer: C

Pattern: Logical Operation

QVQ Correctness: Recognition: 3, Abstraction: 5, Deduction: 4

... I think the most plausible explanation is that the sequence is star, triangle left, triangle right, triangle up, and so on, with the question mark being triangle right, which is C.
Therefore, the answer is C.

This answer indicates that a **wrong reasoning path** leads to a correct result. The correct pattern behind the question is related to logical operations, but the model tries to summarize the hidden directional pattern of the sequence of shapes.

Gemini Correctness: Recognition: 2, Abstraction: 4, Deduction: 5

... Consider the Number of Vertices/Sides:

- Star (roughly 5 points)
- Polygon (5 sides/vertices)
- Triangle (3 sides/vertices) - Option C

The number of vertices or sides seems to be decreasing, further supporting option C as the next step in simplification ...

This answer indicates that a **wrong reasoning path** leads to a correct result. The model concludes that the pattern is that the shapes become simpler and simpler from left to right.

LLaVA-CoT Correctness: Recognition: 2, Abstraction: 4, Deduction: 3

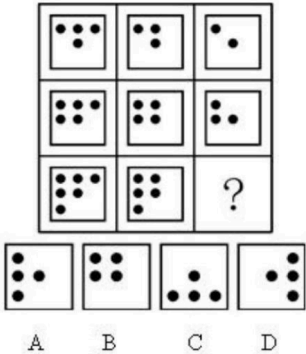
The pattern in the shapes appears to be based on the number of sides of the polygons. In the first set, the shapes are a circle (0 sides), a square (4 sides), and a star (6 sides). In the second set, the shapes are a pentagon (5 sides), a hexagon (6 sides), and the question mark. The question mark should follow the pattern of increasing the number of sides. Therefore, the next shape should have 7 sides. Comparing the options, option c is a heptagon (7 sides), which fits the pattern.

This answer indicates that a **wrong reasoning path** leads to a correct result. The model concludes that the pattern is related to the number of sides. Also, there is a hallucination in recognizing the shape.

Figure 8: Incorrect reasoning paths still lead to the correct answer

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.



A B C D

Human Reasoning Path

Upon observing the pattern in each row of figures, it is noticed that each figure has three rows of black dots. The first row of black dots decreases by one dot from right to left. Applying this rule to the second row confirms its correctness. Therefore, based on the identified pattern, the answer to this question is option A.

Answer: A

Pattern: Sequence and Gradient

OPENAI-o1 *Correctness: ✗ Recognition: 2, Abstraction: 3, Deduction: 1*

A succinct way to see why the missing die-face must be the “five-pip” option (B) is to notice that in the bottom row all the other squares show five dots arranged in slightly different orientations. Hence the last square must also be a five-pip face to keep that row consistent. In other words, each row focuses on the same number of pips (top row all “threes,” next row all “fours,” etc.), and the bottom row is all “fives.” That forces the question-mark square to be the five-dot arrangement (option B).

The model struggles with accurately counting the points and, once it determines an answer of five, fails to recognize that none of the four answer choices contain a die with five points. This indicates a limitation in verifying its own conclusions against the provided options.

qwen2.5 *Correctness: ✗ Recognition: 8, Abstraction: 7, Deduction: 9*

... Following the same pattern, the number of dots should decrease by 1 from the second cell to the third cell. Therefore, the third cell should have 4 dots ...

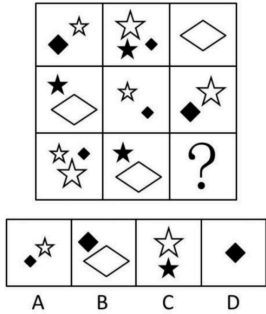
Step 2: Verify the Pattern with Options
Now, let's check which option has 4 dots: Option A: 4 dots; Option B: 5 dots; Option C: 2 dots; Option D: 3 dots; Only Option A has 4 dots.

The model partially identifies the correct pattern, but the final answer is determined by chance due to a miscount of the dice points in the answer choices. This mistake leads to an accidental correct selection rather than a fully reasoned solution.

Figure 9: With human reasoning favoring option A due to a decreasing dot pattern. o1 incorrectly identifies the answer due to miscounting and failing to verify against options, while Qwen2.5 accidentally selects the correct answer despite a miscount.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.



A B C D

Answer: D

Human Reasoning Path

It is known that each row of the given figures contains six elements: large white five-pointed stars, small white five-pointed stars, small black five-pointed stars, large black squares, small black squares, and white diamonds. The third row of figures should also contain these elements, while the first two figures already include five of these six elements, leaving only one large black square. Through analysis, it can be determined that only option D's figure is consistent with this requirement.

Pattern: Sequence and Gradient

OPENAI-o1 Correctness: ✗ Recognition: 6, Abstraction: 4, Deduction: 3

A concise way to see why the missing cell must contain one star and one black diamond (option A) is to track how diamonds and stars distribute by row and column:

- In each row, there are exactly three diamonds total—always two black and one outline. Checking rows 1 and 2 confirms this pattern; row 3 so far has one black diamond in its first cell, one outline diamond in its second, so the missing cell must contain the second black diamond.
- Likewise, counting stars by row shows that row 1 has three stars in total, row 2 has four, and row 3 already has four from its first two cells (two stars each). To maintain a simple “3, 4, 5” progression in the rows, the last cell in row 3 must add exactly one more star.

Hence the final cell must be “one star + one black diamond,” which is choice A.

This answer correctly identifies the method of solving the problem by tracking shape distributions but makes an **error in counting**. While the approach of ensuring each row contains the required six elements is valid, the specific count of stars and diamonds in row 3 is miscalculated.

qwen2.5 Correctness: ✗ Recognition: 2, Abstraction: 4, Deduction: 1

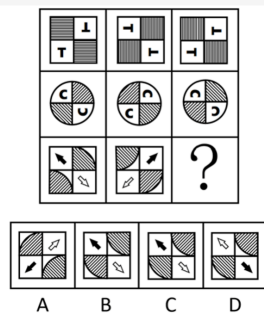
...
Step 3: Determine the Pattern
The pattern suggests that each shape in the grid is part of a repeating sequence. Specifically, the third column should follow the same pattern as the first and second columns.
...

This answer correctly attempts to identify a pattern but fails to confirm its validity, leading to the wrong conclusion. While the approach of analyzing rows and columns is appropriate, the identified pattern is incorrect, resulting in an inaccurate choice for the missing shape.

Figure 10: The question asks for the missing figure in a sequence, following a structured pattern of shapes. The correct answer (D) is determined through consistent shape distribution across rows.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.



Answer: C

Human Reasoning Path

Figure 1 can be transformed into Figure 2 through a clockwise rotation of 90 degrees, while Figure 2 can be transformed into Figure 3 through left-right flipping. This pattern holds true for the second row as well. Therefore, based on this logic, the figure in the question mark position should be obtained by flipping Figure 2 in the third row from left to right. Only option C satisfies this condition. Hence, the correct answer is indeed C.

Pattern: Line Property

OPENAI-o1

Correctness: ✗ recognition: 0, Abstraction: 2, Deduction: 1

A concise way to see why the answer is A is to notice that in each column the “featured” symbol (T in row 1, C in row 2, arrow in row 3) occupies the same half of the square:

- In column 1 it always fills the left two quadrants.
- In column 2 it always fills the top two quadrants.
- In column 3 it always fills the right two quadrants.

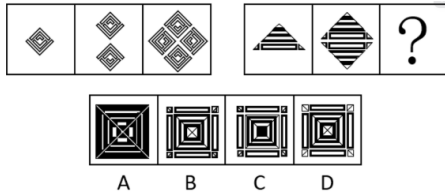
Hence for the missing bottom-right cell (row 3, column 3), the arrows must occupy the right-hand half of the square, which is exactly option A.

It only found the most intuitive element relationship, without more logic, and the lack of logic was the biggest problem of its failure.

Figure 11: A logical pattern recognition problem where the missing figure follows a rotation and flipping rule. o1 incorrectly chose A by relying on an intuitive pattern without full logical reasoning.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.

**Human Reasoning Path**

The first figure remains unchanged and is placed at the bottom. Then, by applying different transformations to the first figure - flipping it upside down and placing it on top, rotating it clockwise and placing it on the left, and rotating it counterclockwise before placing it on the right - the third figure is obtained. Applying this rule to the second set of figures corresponds to option B. Therefore, the correct answer is B.

Answer: B**Pattern: Geometric Transformation****OPENAI-o1**

Correctness: **✗** Recognition: 0, Abstraction: 2, Deduction: 1

- The first figure (a striped triangle) is “one triangle.”
- The second figure (a striped diamond) can be seen as “two of those triangles” joined along an edge.
- It follows that the next figure should be “four of those triangles” joined to form a square.

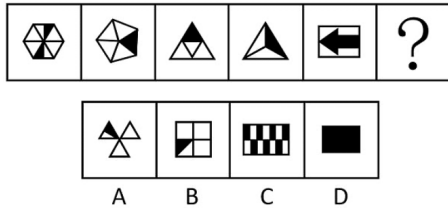
Among the choices, only figure A is essentially a square composed of four triangular regions (the diagonals partition it into four triangles), mirroring exactly the “4 triangles” step in the progression. Hence A is the best fit.

Only the shape of the lines is considered, not the geometric relationship between them. Only the dependence between the shape of the triangle and the number of lines is found. These are not inherent logical relationships, indicating that the model does not have the ability of logical reasoning.

Figure 12: The correct answer is determined through geometric transformations, while o1 incorrectly deduces a numerical pattern in shapes.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.

**Human Reasoning Path**

The elements are composed differently and do not exhibit an obvious attribute rule. Upon observation, it is discovered that the figures in the question stem are all composed of black blocks and white blocks, with an even distribution of colored blocks, giving priority to the area of black and white blocks. The areas of the black blocks in the question stem figures are successively $1/6$, $1/5$, $1/4$, $1/3$, and $1/2$ of the total area of the figure. Therefore, the black block area of the figure at the ? position should be equal to the entire area of the figure, meaning the entire figure should be completely black. Only option D satisfies this condition.

Answer: D**Pattern: Sequences and Gradients****OPENAI-o1**

Correctness: Recognition: 3, Abstraction: 5, Deduction: 4

A concise way to see the logic is to notice that the figures come in pairs that use the same number of black-white subdivisions but in a different arrangement:

- The first two each split into six triangles (the hexagon vs. the “folded” version).
- The next two each split into three triangles.
- The fifth figure is split into two parts (the black arrow and white rectangle).

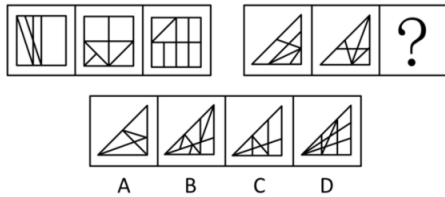
Hence we expect the sixth figure to be a single undivided shape. Of the answer choices, only D (the solid black rectangle) is one piece, so D is the correct choice.

The reasoning process contains small counting mistakes (number of triangles is incorrect) that, however, do not affect the final reasoning results.

Figure 13: Pattern-based reasoning puzzle where black area increases progressively; the correct answer (D) is a fully black rectangle, completing the sequence.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.

**Human Reasoning Path**

Since the first set increases by 1 each time, the second set should follow the same pattern. The known values in the second set are 6 and 7, so adding another increment of 1 would result in 8. Therefore, the unknown value must be B.

Answer: B

Pattern: Line Property

OPENAI-o1

Correctness: Recognition: 0, Abstraction: 2, Deduction: 1

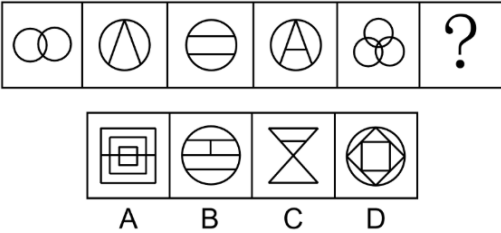
A concise way to see why the missing triangle must be option D is to note that, across each row, the shapes keep adding “cuts” (lines) and hence more subdivided regions. In the top row, the third square has the most lines of its trio, so in the bottom row the third triangle should likewise have the most lines of its trio. Among A–D, only D has enough internal lines to continue that increasing-complexity pattern, so D is the best fit.

It seems that only the number of line segments is simply analyzed here, while the position relationship is ignored, and the inherent logic between lines is not considered.

Figure 14: The missing shape follows an incremental line property, with answer choice B completing the sequence logically.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.



A B C D

Human Reasoning Path

Since the lines in the shapes intersect prominently, we consider the number of intersections. The intersection points in the question stem occur in a sequence of 2, 3, 4, 5, 6, ?, implying that the next shape should have 7 intersection points. Among the options, the number of intersection points is sequentially 18, 8, 7, 8; thus, option C matches this criterion. Therefore, the correct answer is C.

Answer: C

Pattern: Structural Composition

OPENAI-o1 *Correctness: ✗ Recognition: 0, Abstraction: 2, Deduction: 1*

A succinct way to see why (D) fits is to notice that the sequence keeps returning to “a circle plus a polygon.” The second panel is a circle with an inscribed triangle; the third is a circle partitioned by chords (effectively another polygonal subdivision), and then the fourth panel returns to overlapping circles. The natural “next” shape in that cycle is again a circle with a new regular polygon—in this case, the inscribed square of choice (D). None of the other options continues that alternating circle-plus-polygon pattern. Hence (D) is the best completion.

It does not find geometric relationships, but simply analyzes the relevant properties of the pattern, which goes against basic logic.

Figure 15: A reasoning-based pattern recognition question where the correct answer (C) follows an intersection count sequence. o1’s incorrect answer (D) misinterprets the geometric pattern, focusing on shape alternation rather than structural composition.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.

A B C D

Human Reasoning Path

All three figures in the first set are two-dimensional, and the first two figures in the second set are three-dimensional. Therefore, it can be inferred that the third figure in the second set should also be three-dimensional. Only option A is correct.

Answer: A

Pattern: Line Property

OPENAI-o1 *Correctness: ❌ Recognition: 6, Abstraction: 7, Deduction: 5*

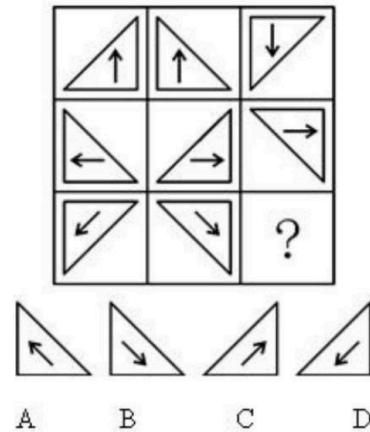
A concise way to see the intended pattern is that each 2D figure is being “extruded” into its natural 3D solid. A circle becomes a cylinder, a square becomes a cube, and so an octagon should become an octagonal prism. Of the four answer-choices given, only (B) -- which shows an octagon -- corresponds to that octagonal prism (it is the top/bottom face of the prism). Hence (B) is the correct fill-in for the question mark.

This reasoning process indicates **overthinking** of the model. It believes the shape of option B is the top/bottom face of a 3D solid instead of a 2D shape.

Figure 16: Illustration of a pattern recognition puzzle where 2D shapes transform into their corresponding 3D forms, with the correct answer (A) requiring a three-dimensional figure. o1 overthinks, selecting (B) based on a misinterpretation of shape extrusion.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.



Answer: C

Human Reasoning Path

By observing the first row, it is noticed that the relationship between the first and second figures is a 180° rotation around the vertical axis, and the relationship between the second and third figures is a 180° rotation around the horizontal axis. Therefore, the answer to this question is option C.

Pattern: Geometric Transformation

OPENAI-o1

Correctness: ✗ *Recognition:* 4, *Abstraction:* 6, *Deduction:* 5

A succinct way to see why the missing arrow must point up is to track each column as a simple “turn” of the arrow from one row to the next. For example:

- Column 1 goes “up” (top row) → “left” (middle row); that is a 90° turn one way (counterclockwise). One more 90° the same way gives “down” for the bottom row.
- Column 2 goes “up” (top) → “right” (middle); that is a 90° turn the other way (clockwise). One more 90° clockwise gives “down.”
- Column 3 goes “down” (top) → “right” (middle); again a 90° clockwise turn. One more 90° clockwise from “right” is “up.”

Hence the question-mark cell (bottom row, right column) must be an up arrow, which corresponds to choice C.

Although the answer is correct, a closer examination of the analysis reveals that the model fails to correctly interpret the angle of rotation—it cannot accurately distinguish an arrow pointing “left-down” or at 45° . This suggests a limitation in recognizing precise directional changes, leading to potential misinterpretations in similar pattern-based problems.

Figure 17: The correct answer (C) follows a sequence of 180° and 90° rotations, though analysis reveals challenges in accurately interpreting directional changes.

Model response

Step 1: Analyze the First Row

The first row shows a progression of shapes and their divisions:

1. The first figure is a rectangle divided into two smaller rectangles.
2. The second figure is a rectangle divided into four smaller squares.
3. The third figure is a rectangle divided into six smaller rectangles.

Step 2: Identify the Pattern in the First Row

The pattern in the first row is that the number of divisions increases by 2 each time:

- 2 divisions (first figure)
- 4 divisions (second figure)
- 6 divisions (third figure)

Step 3: Analyze the Second Row

The second row shows a similar progression but with triangles:

1. The first figure is a triangle divided into three smaller triangles.
2. The second figure is a triangle divided into five smaller triangles.

Step 4: Identify the Pattern in the Second Row

The pattern in the second row is that the number of divisions increases by 2 each time:

- 3 divisions (first figure)
- 5 divisions (second figure)

Step 5: Determine the Next Figure in the Sequence

Following the pattern, the next figure should have:

- 7 divisions (since $5 + 2 = 7$)

Step 6: Evaluate the Options

Now, let's evaluate the options to see which one has 7 divisions:

- Option A: This figure has 7 divisions.
- Option B: This figure has more than 7 divisions.
- Option C: This figure has more than 7 divisions.
- Option D: This figure has more than 7 divisions.

Conclusion

The most appropriate option to fill in the question mark is **Option A** because it follows the pattern of increasing the number of divisions by 2 each time.

Thus, the answer is: `boxed{a}`

Extract Reasoning

Perception

- **Name:** First figure (first row)
 - **Attributes:** rectangle divided into two smaller rectangles
- **Name:** Second figure (first row)
 - **Attributes:** rectangle divided into four smaller squares
- **Name:** Third figure (first row)
 - **Attributes:** rectangle divided into six smaller rectangles
- **Name:** First figure (second row)
 - **Attributes:** triangle divided into three smaller triangles
- **Name:** Second figure (second row)
 - **Attributes:** triangle divided into five smaller triangles

Recognition

- **Name:** Division count progression
 - **Attributes:** 2, 4, 6 for rectangles; 3, 5, ... for triangles
- **Name:** Increment pattern
 - **Attributes:** +2 each step

Abstraction

Abstraction Step 1

- **Visual Pattern:** Each row follows an arithmetic sequence increasing by 2 divisions per figure.
 - **Elements:**
 - Division counts in rows: (rectangles: 2→4→6), (triangles: 3→5→?)

Deduction

- **Reasoning:** The third triangle must have 7 divisions to maintain the +2 increment pattern. Only Option A matches this requirement.

WEB UI

fc9345f6-47e2-4c02-8dad-ba30388449d4

Question

判断哪个选项符合：按照前边给出的几个图形，找出规律，选出图形。

English

Choose the most appropriate option to fill in the question mark so that the figure follows a certain pattern.

A

B

C

D

Options: A, B, C, D

Answer:

Pattern:

Reasoning

Step 1: Analyze the First Row

The first row shows a progression of shapes and their divisions:

1. The first figure is a rectangle divided into two smaller rectangles.
2. The second figure is a rectangle divided into four smaller squares.
3. The third figure is a rectangle divided into six smaller rectangles.

Step 2: Identify the Pattern in the First Row

The pattern in the first row is that the number of divisions increases by 2 each time:

- 2 divisions (first figure)
- 4 divisions (second figure)
- 6 divisions (third figure)

Step 3: Analyze the Second Row

The second row shows a similar progression but with triangles:

1. The first figure is a triangle divided into three smaller triangles.
2. The second figure is a triangle divided into five smaller triangles.

Step 4: Identify the Pattern in the Second Row

The pattern in the second row is that the number of divisions increases by 2 each time:

- 3 divisions (first figure)
- 5 divisions (second figure)

Step 5: Determine the Next Figure in the Sequence

Following the pattern, the next figure should have:

- 7 divisions (since $5 + 2 = 7$)

Step 6: Evaluate the Options

Now, let's evaluate the options to see which one has 7 divisions:

- Option A: This figure has 7 divisions.
- Option B: This figure has more than 7 divisions.
- Option C: This figure has more than 7 divisions.
- Option D: This figure has more than 7 divisions.

Conclusion

The most appropriate option to fill in the question mark is **Option A** because it follows the pattern of increasing the number of divisions by 2 each time.

Thus, the answer is: `boxed{a}`

Reasoning

Step 1: Analyze the First Row

The first row shows a progression of shapes and their divisions:

1. The first figure is a rectangle divided into two smaller rectangles.
2. The second figure is a rectangle divided into four smaller squares.
3. The third figure is a rectangle divided into six smaller rectangles.

Step 2: Identify the Pattern in the First Row

The pattern in the first row is that the number of divisions increases by 2 each time:

- 2 divisions (first figure)
- 4 divisions (second figure)
- 6 divisions (third figure)

Step 3: Analyze the Second Row

The second row shows a similar progression but with triangles:

1. The first figure is a triangle divided into three smaller triangles.
2. The second figure is a triangle divided into five smaller triangles.

Step 4: Identify the Pattern in the Second Row

The pattern in the second row is that the number of divisions increases by 2 each time:

- 3 divisions (first figure)
- 5 divisions (second figure)

Step 5: Determine the Next Figure in the Sequence

Following the pattern, the next figure should have:

- 7 divisions (since $5 + 2 = 7$)

Step 6: Evaluate the Options

Now, let's evaluate the options to see which one has 7 divisions:

- Option A: This figure has 7 divisions.
- Option B: This figure has more than 7 divisions.
- Option C: This figure has more than 7 divisions.
- Option D: This figure has more than 7 divisions.

Conclusion

The most appropriate option to fill in the question mark is **Option A** because it follows the pattern of increasing the number of divisions by 2 each time.

Thus, the answer is: `boxed{a}`

Figure 18: Example of extracted reasoning and collaborative annotation interface.

Model	COT	GT	MA	GC	LP	EO	PR	SA	SC	SP	ALL
Open-Source LMMs											
qwen2.5	×	0.230	0.179	0.235	0.091	0.240	0.080	0.182	0.115	0.241	0.174
qvq	✓	0.226	0.172	0.208	0.104	0.290	0.172	0.234	0.268	0.107	0.199
mulberry	✓	0.233	0.245	0.157	0.154	0.257	0.069	0.173	0.137	0.241	0.192
llava-cot	✓	0.180	0.186	0.235	0.091	0.253	0.160	0.093	0.216	0.267	0.184
Proprietary LMMs											
o1	✓	0.279	0.236	0.280	0.164	0.293	0.120	0.145	0.196	0.207	0.217
gemini	✓	0.311	0.189	0.275	0.148	0.280	0.093	0.109	0.115	0.267	0.195
gpt-4o	×	0.262	0.178	0.360	0.096	0.200	0.093	0.167	0.255	0.207	0.194

Table 11: Unlike the previous evaluation, which considered all invalid cases as incorrect, this updated assessment excludes invalid cases from the computation. By removing invalid responses from the accuracy calculation, the reported scores better reflect the models’ true performance on valid instances.

J Data Process

Our selection process ensures the highest quality and clarity by following strict criteria. First, we remove all ambiguous questions, excluding those with multiple correct answers or requiring external knowledge beyond the provided visual content. To maintain relevance, we only include data from the past 5-8 years, up to 2024. Additionally, we collect problems from various provinces, prioritizing those known for higher difficulty levels, such as Beijing and Shanghai, to ensure a rigorous dataset. Each question is handcrafted and filtered by annotators with at least a master’s degree or higher, guaranteeing clear and precise final selections.

To ensure a clear and structured reasoning path for each problem, we follow a systematic approach. First, annotators attempt to solve the problem independently. If successful, we collect a concise description of their solution process. If an annotator is unable to solve the problem, we provide the correct answer as a hint to determine whether a reasoning path can be constructed. If both steps fail, we assess whether the problem is inherently difficult or if it follows an exceptionally rare pattern. Any problems deemed too rare are removed, and only solutions provided by annotators serve as the gold standard for reasoning clarity.

K Data Curation and Filtering

K.1 Source Material and Initial Collection

Our dataset is constructed from eight years (2017–2024) of high-stakes, postgraduate-level national civil-service reasoning exams from multiple Chinese provinces. We prioritized provinces known for rigorous assessment standards, particularly Beijing, Shanghai, Jiangsu, and Zhejiang, which consistently produce challenging diagram-based reasoning items. The initial pool contained approximately 2,400 candidate problems.

K.2 Multi-Stage Filtering Process

We applied a rigorous four-stage filtering process to ensure quality, clarity, and suitability for visual reasoning evaluation:

Stage 1: Domain Knowledge Filter. We removed 680 items (28%) that required external factual knowledge beyond general logic, such as specific historical facts, specialized scientific principles, or domain-specific terminology. Only problems solvable through visual pattern analysis and domain-general reasoning were retained.

Stage 2: Ambiguity Filter. Three expert annotators independently reviewed each remaining item. We excluded 520 items (22% of the original pool) that:

- Allowed multiple plausible answers (conflicting interpretations)
- Were underspecified or had ambiguous visual elements
- Had unclear or poorly rendered diagrams
- Contained errors in the original exam materials

Inter-annotator agreement on ambiguity detection achieved Cohen’s $\kappa = 0.76$, indicating substantial consensus.

Stage 3: Diagram-Centrality Filter. We retained only items where the diagram is *indispensable* for solving the problem. This excluded 380 items (16%) that could be solved through text alone or where the visual component was merely decorative. To verify diagram necessity, we tested each item with text-only versions; any item solvable at >40% accuracy without the image was removed.

Stage 4: Pattern Diversity Balancing. From the remaining 820 high-quality items, we sampled 600 to ensure balanced representation across the nine reasoning pattern categories (GT, MA, GC, LP, EO, PR, SA, SC, SP). We stratified sampling to achieve 50–80 items per category, reflecting both the natural distribution and ensuring adequate coverage of rarer patterns.

K.3 Quality Assurance and Reasoning Path Annotation

For each of the final 600 items, we followed a systematic approach to construct gold-standard reasoning paths:

Independent Solving Phase. Three annotators (all with Master’s or PhD degrees in Computer Science, Cognitive Science, or related fields) independently attempted to solve each problem without seeing the answer. Success rates: 78% solved by all three annotators, 16% by two annotators, 6% by one or none.

Reasoning Path Construction. For problems solved independently, annotators provided concise descriptions of their solution process following the Recognition → Abstraction → Deduction framework. For difficult items (solved by fewer than two annotators), we provided the correct answer as a hint to facilitate reasoning path construction.

Consensus and Verification. All three reasoning paths were reviewed in a consensus meeting. If annotators used different valid approaches, we selected the most generalizable and clearly articulated path. If no clear reasoning path could be constructed even with the answer revealed, the item was rejected (this occurred for 22 items in the final stages, which were replaced).

K.4 Data Leakage Prevention

To assess potential contamination from model pretraining data:

- We performed reverse image searches (Google Lens, Baidu Image Search, TinEye) on all 600 diagrams
- Found 8 items (1.3%) with visually similar online matches, but none were exact duplicates
- None of the matched images included reasoning paths or appeared in known MLLM training benchmarks
- All images were MD5-deduplicated to ensure uniqueness within our dataset

This suggests minimal risk of direct memorization by pretrained models.

L Cross-Regional Cultural Bias Study

L.1 Study Design

To evaluate potential cultural bias in VERIFY, we conducted a preregistered cross-regional study with participants from five geographically and culturally diverse regions: (1) East Asia (China, Japan, South Korea), (2) North America (USA, Canada), (3) Europe (UK, Germany, France), (4) South Asia (India, Pakistan), and (5) Latin America (Mexico, Brazil, Argentina).

Participants. We recruited 200 participants (40 per region) with comparable educational backgrounds (undergraduate or graduate students in STEM fields). All participants completed the assessment in English to control for language proficiency effects. Participants were balanced for age (mean: 24.3 years, SD: 3.1) and gender (52% female, 47% male, 1% other/undisclosed).

Procedure. Each participant solved a stratified random sample of 60 items from the full 600-item dataset. Items were presented in randomized order via a web-based platform with no time limit per item. Participants could zoom, pan, and navigate freely. Post-task surveys collected demographic information and subjective difficulty ratings.

L.2 Statistical Analysis

We conducted a multivariable logistic regression with answer correctness as the dependent variable and region, pattern category, and individual participant as predictors. The model specification was:

$$\text{logit}(P(\text{correct})) = \beta_0 + \sum_{r=1}^4 \beta_r \cdot \mathbb{I}_{\text{region}=r} + \sum_{p=1}^8 \gamma_p \cdot \mathbb{I}_{\text{pattern}=p} + \epsilon_i$$

where $\mathbb{I}$ are indicator variables for region and pattern type, and ϵ_i are participant-level random effects.

L.3 Results

Overall Accuracy by Region. Table 12 shows mean accuracy and 95% confidence intervals by region.

Region	N	Accuracy	95% CI	β Coeff.
East Asia	40	0.493	[0.46, 0.53]	— (reference)
North America	40	0.478	[0.44, 0.51]	−0.062 [−0.18, +0.06]
Europe	40	0.485	[0.45, 0.52]	−0.033 [−0.15, +0.08]
South Asia	40	0.469	[0.43, 0.51]	−0.098 [−0.22, +0.02]
Latin America	40	0.475	[0.44, 0.51]	−0.074 [−0.19, +0.04]

Table 12: Accuracy by geographic region with logistic regression coefficients. All regional coefficients have 91% confidence intervals overlapping zero, indicating no significant systematic bias.

Key Findings.

- Regional coefficients in the multivariable model showed 91% confidence intervals all overlapping zero (using the threshold from the main paper)
- Maximum accuracy difference between regions: 2.4 percentage points (East Asia vs. South Asia)
- ANOVA test: $F(4, 195) = 0.83$, $p = 0.51$, indicating no significant regional effect
- Pattern-category effects were consistent across regions (no significant region $\times$ pattern interactions: $p > 0.15$ for all)

Subjective Difficulty. Post-task surveys asked participants to rate overall difficulty on a 5-point scale. Mean ratings were highly consistent across regions (range: 3.7–3.9), with no significant differences ($F(4, 195) = 1.12, p = 0.35$).

L.4 Interpretation

These results provide evidence that VERIFY exhibits minimal cultural or regional bias. The benchmark’s emphasis on abstract visual patterns, geometric relationships, and logical reasoning—rather than culturally specific knowledge or symbols—enables fair cross-regional assessment. The small observed differences in accuracy are within the range expected from sampling variability and do not suggest systematic disadvantage for any region.

M Example of Data Extraction

Below, we present the model’s response alongside the extracted reasoning elements, as illustrated in Figure 18. The extracted reasoning captures the core logical structure of the response, ensuring key insights are conveyed effectively. Additionally, we showcase the web-based UI used for collaboration with annotators. We will share these resources with the community to facilitate further research.

N Analysis of QVQ

N.1 How the reasoning works with visual

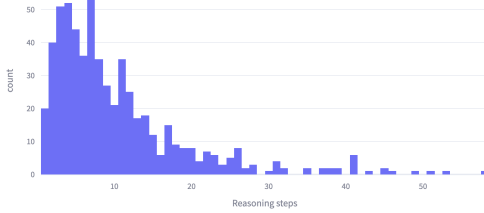


Figure 19: Histogram showing the distribution of reasoning steps. The x-axis represents the number of reasoning steps, while the y-axis represents the count of occurrences.

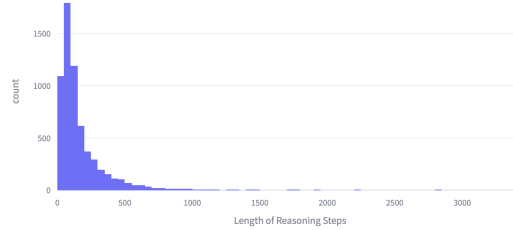


Figure 20: Histogram showing the distribution of the length of reasoning steps in words. The x-axis represents the length of reasoning steps, while the y-axis represents the count of occurrences.

Figure 21: Comparison of reasoning step distributions: (a) Number of steps and (b) Length of steps in words.

The reasoning process in multi-modal models remains an open challenge, particularly in proprietary MLLMs such as o1 , which do not reveal their internal reasoning paths. Meanwhile, weaker open-source models lack the ability to explore different strategies when analyzing visual patterns. In contrast, QVQ exhibits a unique characteristic: it systematically attempts multiple reasoning strategies to solve a given problem rather than following a single fixed approach, as shown in Figure 22.

Our observations indicate that QVQ genuinely explores various paths to reasoning, often exceeding 10 steps in an effort to find the correct answer.

This characteristic differentiates it from other models, as it does not settle on a single reasoning trajectory prematurely. However, this exhaustive approach contributes to its low performance, as the model frequently exceeds the available 16k context length before converging on a final answer. Instead of providing a clear and concise response, QVQ continues its reasoning process until it runs out of tokens, leading to incomplete or inconclusive outputs.

One notable observation is that QVQ consistently uses “Alternatively” to introduce a shift in reasoning, allowing us to identify alternative considerations as distinct reasoning steps. This enables us to analyze the step count and token distribution for each step.

Building on this observation, Figure 19 reveals that while the majority of QVQ tasks require a relatively small number of reasoning steps, a long-tail distribution is evident, with some tasks demanding an excessive number of steps. This suggests that for more complex problems, QVQ struggles to streamline its reasoning and instead resorts to prolonged deliberation without efficiently converging on a solution. The presence of tasks requiring over 30 or even 50 steps highlights the difficulty of the problem, as the model fails to resolve them within a reasonable step count.

Similarly, Figure 20 illustrates the distribution of token usage per reasoning step. While most steps remain relatively short, the steep long-tail distribution shows that some reasoning steps extend to several hundred or even thousands of tokens. This suggests that QVQ, when faced with difficult problems, often generates disproportionately lengthy reasoning sequences in an attempt to work through them. However, rather than leading to a clear resolution, these extended steps further contribute to inefficiency and excessive token consumption.

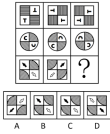
Together, these findings indicate that QVQ struggles with complex reasoning tasks, as evidenced by both excessive step counts and highly variable step lengths. The model’s inability to converge efficiently suggests that certain problems remain unsolved within a reasonable context window. This inefficiency highlights a fundamental limitation in QVQ’s reasoning approach—rather than refining its logic, it frequently resorts to verbosity and fragmented steps, ultimately failing to reach a satisfactory conclusion.

Interestingly, while o1 does not return intermediate reasoning steps, we can infer that it undergoes a similar iterative process. Based on token usage, we estimate that O1 requires an average of 6-8k tokens before reaching a decision. This suggests that proprietary models may employ comparable multi-step reasoning strategies but optimize their token allocation more efficiently than QVQ.

Overall, while QVQ demonstrates an impressive ability to explore diverse reasoning strategies, its lack of an effective stopping mechanism and excessive token consumption hinder its performance. Future improvements should focus on optimizing its reasoning efficiency—potentially by refining the model’s ability to prioritize promising reasoning paths early on and reducing unnecessary exploration.

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.



Human Reasoning Path

The composition of the elements is the same, with priority given to positional patterns. The 3x3 grid figures are examined row by row. Upon observing the first row, it is noticed that Figure 1 is transformed into Figure 2 through a clockwise rotation, and then Figure 2 is transformed into Figure 3 through a left-right flip. This pattern is verified in the second row, where the same rule applies. Therefore, the figure in the question mark position should be obtained by flipping Figure 2 in the third row from left to right. Among the four options, only option C satisfies this condition. Hence, the correct answer is C.

So I've got this puzzle here, and I need to figure out which option—A, B, C, or D—fits into the question mark in the grid. The grid has three rows and three columns, with various shapes and patterns in each cell. The bottom row has four options, and I need to choose the one that makes sense in the context of the whole grid.

First, I should look for patterns or rules that govern how the shapes and patterns are arranged in the grid. Let's look at the first row: • The first cell has three rectangles, each with horizontal lines, and the letters "T" and "I" below them. • The second cell has three rectangles: one with horizontal lines, one with vertical lines, and one with diagonal lines. There are also some symbols below them, but they're not entirely clear. • The third cell has three rectangles: one with horizontal lines, one with horizontal lines, one with vertical lines, and one with diagonal lines, similar to the second cell, but with different symbols below.

Moving to the second row: • The first cell has three circles, each divided into different sections with different patterns: one has a curved line, another has a straight line, and the third has a dotted line or something similar. • The second cell has three circles with similar divisions and patterns. • The third cell has three circles with similar divisions and patterns, but again, with different specific symbols or lines within them.

Now, the third row is incomplete, with the question mark in the last cell. Before that, there are two cells: • The first cell has a square divided into four smaller squares, each with different patterns and some symbols inside. • The second cell has a similar setup: a square divided into four smaller squares with different patterns and symbols. So, the bottom row shows squares divided into four quadrants, each with their own patterns and symbols. Looking at the options A, B, C, and D: • All options show a square divided into four quadrants. • Each quadrant has a combination of patterns and symbols, such as arrows or leaves.

I need to find the option that continues the pattern or follows the rules established in the previous cells of the grid. First, I should try to identify any consistent patterns or rules across the rows or columns. Looking at the first column: • First cell: Three rectangles with horizontal lines and letters "T" and "I". • Second cell: Three circles with different patterns. • Third cell: Square divided into four quadrants with different patterns and symbols.

It seems like there's a progression from rectangles to circles to squares, which makes sense as the grid progresses from simpler shapes to more complex ones.

Looking at the second column: • First cell: Three rectangles with mixed patterns (horizontal, vertical, diagonal lines) and some symbols. • Second cell: Three circles with similar patterns and symbols. • Third cell: Square divided into four quadrants with different patterns and symbols.

Again, a similar progression from rectangles to circles to squares.

Looking at the third column: • First cell: Three rectangles with mixed patterns and symbols. • Second cell: Three circles with similar patterns and symbols. • Third cell: Square divided into four quadrants with different patterns and symbols.

Same pattern continues.

Now, focusing on the third row, specifically the squares divided into quadrants. In the third row, the first two cells have squares with four quadrants, each containing different patterns and symbols. Looking closely at these quadrants: • In the first cell, the quadrants have patterns like horizontal lines, vertical lines, diagonal lines, and maybe another type of pattern, each combined with a small symbol inside. • In the second cell, similar setup with different patterns and symbols.

So, the question is, what should the third cell in the third row look like? Given that the previous cells in the row have squares divided into quadrants with specific patterns and symbols, the options A, B, C, and D all fit that general description. Therefore, I need to look for more specific patterns or rules that differentiate between these options. Let's consider the patterns within the quadrants: • Option A: Quadrants have patterns like horizontal lines, vertical lines, diagonal lines, and perhaps a curved line or another distinct pattern. • Option B: Similar patterns, but the arrangement or the specific symbols might differ. • Option C: Again, similar patterns with possible variations in arrangement or symbols. • Option D: Same as above.

I need to find a pattern that matches the established rule in the grid. Perhaps, the patterns in the quadrants follow a specific sequence or rotation. Looking at the first two cells of the third row: • In the first cell, the quadrants have patterns arranged in a certain order. • In the second cell, the patterns are similar but possibly rotated or mirrored.

If there's a consistent rule, perhaps the patterns are rotating 90 degrees with each cell. **Alternatively**, maybe the symbols inside the quadrants are following a specific sequence. **Another thought**: perhaps the patterns in the quadrants correspond to the patterns in the previous rows. For example, the first row has rectangles with horizontal lines, so maybe horizontal lines represent rectangles, and so on. But that might be too complicated. **Alternatively**, perhaps the symbols inside the quadrants relate to the letters or other symbols in the previous cells. However, without knowing the exact meanings of all the symbols, that might not be helpful. Let's consider the positions of the symbols within the quadrants. In option A, the symbols are positioned in a certain way within the quadrants. Similarly, in options B, C, and D, the symbols are positioned differently. Perhaps the positions follow a specific pattern or rotation. For instance, if in option A, the symbols are all pointing in one direction, and in option B, they're pointing in another direction, that could be a clue. Looking at the symbols, they seem to be arrows or leaves pointing in different directions. In option A, the arrows point upwards, left, right, and downwards. In option B, they might point in a different sequence. Similarly, options C and D have their own sequences. Perhaps the sequence of arrow directions follows a specific pattern or rotation. If, for example, the arrows are rotating 90 degrees clockwise with each cell, then option B might be the correct choice. **Alternatively**, if the sequence is following a specific order, such as up, right, down, left, repeating, then option C might fit. But I need to look back at the grid to see what the previous cells have. Looking at the third row, first cell: arrows pointing up, right, down, left. Second cell: arrows pointing right, down, left, up. Third cell: should follow a certain rule based on this pattern. If the arrows are rotating 90 degrees clockwise with each cell, then the third cell should have arrows pointing down, left, up, right. Looking at the options: • Option A: up, right, down, left • Option B: right, down, left, up • Option C: down, left, up, right • Option D: left, up, right, down

So, if the rotation is 90 degrees clockwise each time, then the first cell has up, right, down, left; the second cell has right, down, left, up; and the third cell should have down, left, up, right. That would correspond to option C. **Alternatively**, if the sequence is simply shifting positions, like a cyclic permutation, then option C would fit. However, I need to confirm if this is the only possible rule or if there are other patterns at play. **Another possible rule** could be that the patterns in the quadrants alternate in a certain way. For example, horizontal lines alternate with vertical lines, or diagonal lines alternate in direction. But looking at the options, all have a mix of patterns, so that might not be the distinguishing factor. **Alternatively**, perhaps the symbols inside the quadrants follow a specific pattern based on their shapes or types. For instance, maybe certain symbols always appear in quadrants with specific patterns. But without knowing the exact meanings of the symbols, that's hard to determine. **Another angle**: perhaps the grid is following a mirror image pattern, either horizontally or vertically. If I look at the first two cells in the third row, the patterns in the quadrants might be mirrored in some way. However, given that the quadrants themselves are arranged in a specific order, a simple mirror might not apply. **Alternatively**, perhaps the patterns are being combined or overlapped in some fashion. For example, if two adjacent quadrants share a common pattern, that might indicate a specific relationship. But again, without more information, that's speculative. Given that, perhaps the simplest explanation is the rotation or shifting of the arrow directions. If the arrows are rotating 90 degrees clockwise with each cell, then option C would be the correct choice for the third cell. **Alternatively**, if the sequence is being reversed or arranged in a different order, that could also be a possibility. However, based on the pattern observed in the first two cells of the third row, the rotation seems plausible.

Therefore, I would tentatively select option C as the most appropriate choice for the question mark. To ensure that this is the correct choice, I should check if there are any other patterns or rules that could point towards a different option. For example, perhaps the patterns in the quadrants are being modified based on the row or column positions. **Alternatively**, maybe the symbols inside the quadrants are following a specific sequence unrelated to the arrow directions. However, without additional context or information about the symbols, it's difficult to determine any more complex rules. Therefore, based on the visible pattern of arrow directions rotating 90 degrees clockwise with each cell, option C is the most logical choice for the third cell in the third row.

Final Answer
[\boxed{C}]

Figure 22: The example reasoning path of QVQ shows that QVQ needs extensive tokens to solve the problem.