# When Visualizing is the First Step to Reasoning: MIRA, a Benchmark for Visual Chain-of-Thought

Supplementary Material



## 6. Experimental Model Settings

This section details the configurations for all models evaluated in our experiments. For all API-based models, we utilized the default decoding settings provided by each endpoint, with the maximum output length set to 16,384 tokens. The specific versions and checkpoints are organized in Table 3.

For a specific subset of models, we used tailored generation parameters. Specifically, for Qwen-VL-Max (325B), GLM-4.5V (106B), and both variants of Qwen2.5-VL (32B/72B), we set the maximum output length to 8,192 tokens and used a top$_p$ value of 1.0. For the Bagel (operating in thinking mode) and Janus-Pro models, we followed the official inference configurations from their respective code repositories to ensure faithful evaluation.

## 7. Prompt Settings

This section provides the specific prompt templates used for the three evaluation levels described in Section 3.2, as well as the specialized templates used for the upgraded Text-CoT analysis in Section 4.3. We also include, at the end, the prompt provided to gpt-4o-2024-11-20 during our evaluation.

**Level 1: Direct Evaluation.** For the direct evaluation setting, a straightforward prompt was used to ask the model for the final answer without requesting intermediate reasoning steps. The template was:

> [Input Image]
> prompt = "'Question: {question}
> Please provide the final answer directly. The final answer is placed in <answer></answer>."'

**Level 2: Text-CoT Reasoning.** This level tested the models' ability to use text-based reasoning on **MIRA** tasks. The model is prompted to first generate a textual chain of thought and then provide the final answer. Two types of templates were employed to investigate the efficacy of this approach:

- **General Template ($T_{gen}$):** This approach used a generic CoT prompt for all tasks. It served as a baseline to measure the general applicability of text-based reasoning.

> [Input Image]
> prompt = "'Question: {question}
> Please first conduct step-by-step reasoning, and then provide the final answer. The final answer is placed in <answer></answer>."'

- **Specialized Template ($T_{spec}$):** To test the upper-bound performance of text-only reasoning, dedicated and task-specific CoT prompt templates were designed for each of the 20 tasks in the MIRA dataset. Below are the specific prompts used for each task, organized by category.

**(1) Euclidean Geometry (EG)**

**Task: Convex Hull**

> [Input Image]
> prompt = "'This is a convex hull problem. Analyze the points and determine the vertices of the convex hull. Question: {question}
> Please reason step-by-step: 1. Start with one color (e.g., Red): - Visually/algorithmically assess which Red points are extreme (cannot be expressed as a convex combination of other points). - Count how many target points this Red convex hull would contain (on the boundary or strictly inside). Note any collinear runs along edges and whether intermediate collinear points should be kept or skipped per the task convention. 2. Switch to the other color (e.g., Blue): - Repeat the same analysis: identify extreme Blue points and count how many target points the Blue convex hull contains. 3. Cross-check and reconcile: - Compare Red- and Blue-based findings; verify no interior point is mistakenly classified as a hull vertex. - Use supporting checks (orientation tests/cross products) to confirm each candidate vertex lies on the outer boundary; handle collinearity consistently (keep only endpoints unless the problem requires listing all boundary points). 4. Construct the final hull: - Order vertices counterclockwise starting

Table 3. A comprehensive list of the models evaluated in our experiments. For all API-based models, the default decoding settings were used, as no specific sampling parameters (e.g., temperature) were set.

| Model | Creator | Version / Checkpoint |
|---|---|---|
| GPT-5 | OpenAI | `gpt-5-2025-08-07` |
| GPT-5.2 | OpenAI | `gpt-5.2-2025-12-11` |
| GPT-5-mini | OpenAI | `gpt-5-mini-2025-08-07` |
| GPT-4.1 | OpenAI | `gpt-4.1-2025-04-14` |
| GPT-4.1-mini | OpenAI | `gpt-4.1-mini-2025-04-14` |
| GPT-4o | OpenAI | `gpt-4o-2024-11-20` |
| GPT-4o-mini | OpenAI | `gpt-4o-mini-2024-07-18` |
| o3 | OpenAI | `o3-2025-04-16` |
| o4-mini | OpenAI | `o4-mini-2025-04-16` |
| Claude 4 Opus | Anthropic | `claude4-opus` |
| Claude 4 Sonnet | Anthropic | `claude4-sonnet` |
| Gemini 2.5 Pro | Google | `gemini-2.5-pro` |
| Gemini 2.5 Flash | Google | `gemini-2.5-flash-preview-05-20` |
| Gemini 3 Pro | Google | `gemini-3-pro-preview` |
| Gemini 3 Flash | Google | `gemini-3-flash-preview` |
| Seed1.6 Vision Pro | ByteDance | `doubao-seed-1.6-vision-250815` |
| Seed1.5-VL | ByteDance | `doubao-1.5-vision-pro-250328` |
| Qwen-VL-Max | Alibaba | `qwen-vl-max-0813` |
| Qwen3-VL (235B) | Alibaba | `Qwen3-VL-235B-A22B-Instruct` |
| Qwen-2.5-VL (32B) | Alibaba | `qwen2.5-vl-32b-instruct` |
| Qwen-2.5-VL (72B) | Alibaba | `qwen2.5-vl-70b-instruct` |
| GLM 4.5V (106B) | ZAI | `glm-4.5v` |

from the leftmost-lowest point (or another clear anchor) and ensure the polygon is simple and closed. - Provide the set/list of hull vertices (by labels or coordinates) and the total count. 5. Briefly justify: - Summarize why each listed vertex is extreme and why excluded points are interior or collinear intermediates.
The final answer is placed in <answer></answer>."'

**Task: Overlap**

[Input Image]
prompt = "'Choose two images from A–D and overlay them by aligning their black coordinate-axis borders. This produces the overlapping region of the two shapes. Which pair has the largest overlapping area? Output only two letters like 'AC'. Please reason step-by-step: 1. Normalize: confirm all four tiles share the same scale and origin; treat overlays as perfect border-to-border alignment with no extra rotation/translation. 2. For each pair (AB, AC, AD, BC, BD, CD): - Compare centers and orientations; note how much their silhouettes intersect (heart/square/star/arrow) when placed at identical coordinates. - Use bounding boxes as a quick upper bound; then refine with edge/vertex relationships to judge whether overlap is large (broad interior intersection), medium (partial edge/vertex overlap), or small (mostly disjoint). 3. Track the estimated overlap area (qualitatively or numerically if obvious from symmetry/containment). Resolve ties by preferring the pair with broader interior overlap rather than thin edge contact. 4. State the chosen pair and a 1–2 sentence justification referencing the relative placements/orientations that cause maximal intersection.
The final answer is placed in <answer></answer>."'

**Task: Localizer**

[Input Image]
prompt = "'Tile the square on the right using the

solid-outlined puzzle pieces on the left. Use all pieces; the tiling must be exact—no leftovers, no gaps, no overlaps. Each piece has a circle. After completing the tiling, return the circle coordinates in numerical order using the format: [pieceID, (x, y)]; separate entries with semicolons. Assumptions: use the same unit grid as shown; coordinates are 1-indexed with (x,y) labeled along the top/left axes; rotations and flips are allowed unless forbidden by outlines. Please reason step-by-step: 1. Parse the target grid: record its outer size (width × height) and axes labels. 2. Catalog each piece (1–4): sketch its unit-square footprint, edge types (axis-aligned vs diagonal), and the circle's offset in the piece's local coordinates. 3. Area & boundary check: verify the sum of piece areas equals the target area; note unique constraints (e.g., long diagonals, notches) that can only fit specific borders/corners. 4. Plan placements: anchor the largest/most constrained piece(s) to borders/-corners first; ensure diagonals match the grid diagonals; avoid creating unreachable cavities. 5. Place all pieces: finalize positions and orientations so the region is fully covered; confirm no overlaps and all borders align with grid lines/-diagonals. 6. Convert circle positions: for each placed piece, transform the circle's local offset to global grid coordinates (x, y) and round to exact grid intersections if applicable. 7. Output strictly in the required order and format: [1, (x1, y1)]; [2, (x2, y2)]; [3, (x3, y3)]; [4, (x4, y4)].
The final answer is placed in <answer></answer>."'

**Task: Mirror Pattern**

[Input Image]
prompt = "'Which option (A–D) can be obtained by mirroring the original image once? You may follow these steps to reason: 1) Horizontally mirror the original image. 2) After the reflection, allow an arbitrary in-plane rotation and critically compare against each option A–D (match landmark positions/orientations; rule out any option that would require a second reflection or non-rigid warping).
The final answer is placed in <answer></answer>."'

**Task: Cubes Count**

[Input Image]
prompt = "'What's the number of cubes presented in the image? Please follow these steps: 1. Identify each layer from bottom to top. 2. For each layer, count how many cubes are present. 3. Add up the counts to get the total number of cubes.
The final answer is placed in <answer></answer>."'

**Task: Cubes Missing**

[Input Image]
prompt = "'What is the number of cubes needed to fill in the structure so that it becomes a solid block with no internal gaps? Please follow these steps: 1. Identify the full dimensions of the solid block (length × width × height). 2. For each layer (from top to bottom), count: - Maximum possible cubes in that layer if solid - Actual cubes present - Missing cubes = (full layer) – (present layer) 3. Add the missing cubes across all layers.
The final answer is placed in <answer></answer>."'

**(2) Physics-Based Reasoning (PBR)**

**Task: Billiards**

[Input Image]
prompt = "'In the image, a billiards table has pockets labeled 1–6. The blue ball rolls along the green arrow, with no spin, perfectly elastic cushion bounces, and unlimited momentum. Which numbered pocket will it finally enter? Answer with a single digit 1–6. You may follow these steps to reason: 1) Normalize the table: record the ball's starting point and the arrow's direction; pockets are fixed at labels 1–6. 2) Use the mirror (unfolding) method: virtually reflect the table across a cushion each time the path would bounce. Extend the initial ray straight through these mirrored copies until it hits the center of a mirrored pocket. 3) Map that hit back to the original table to identify the real pocket number; equivalently, enforce equal-angles for each bounce and verify the same destination. 4) Output only the pocket label (1–6).
The final answer is placed in <answer></answer>."'

**Task: Electric Charge**

[Input Image]
prompt = "'Question: {question}

You may follow these steps to reason:

1. Parse the setup: list each charge with sign, magnitude, and coordinates; identify the target object (which charge/point the net force is asked about).
2. For each source charge, determine the force direction on the target (attraction if opposite sign, repulsion if same sign); sketch/describe the vector qualitatively.
3. Compute each force's magnitude with Coulomb's law

$$|F_i| = k\frac{|q_i q_t|}{r_i^2},$$

and compute vector components using the displacement unit vector from source to target.
4. Apply superposition: sum the components $F_x$ and $F_y$ to obtain the net force vector; use symmetry to simplify whenever possible.
5. Report the net magnitude

$$\sqrt{F_x^2 + F_y^2}$$

and direction (angle or cardinal description), and check limiting/special cases (e.g., $r = 0$ excluded, equal/opposite charges cancel along symmetry axes).

The final answer is placed in <answer></answer>."'

**Task: Mirror Clock**

[Input Image]
prompt = "'Question: {question}

You may reason as follows:

1. First mirror the clock face (by default, a left–right reflection about the vertical axis).
2. Record the hands' angles relative to 12 o'clock after mirroring. The angles transform as

$$\theta' = 360^\circ - \theta,$$

i.e. clockwise and counterclockwise directions swap.
3. If required to match choices/diagram, you may then apply an in-plane rotation (0°, 90°, 180°, 270°), but do *not* perform a second reflection.
4. Convert the mirrored angles back to time and handle hour-minute carry. For minutes $m$ and hours $h$ (12-hour clock, with $h \in \{1, \ldots, 12\}$):

$$m' \equiv (60 - m) \pmod{60},$$

and define the borrow/carry as

$$\text{carry} = \begin{cases} 1, & m \neq 0, \\ 0, & m = 0. \end{cases}$$

Then compute the mirrored hour

$$h' \equiv (12 - h - \text{carry}) \pmod{12}.$$

When presenting the result convert hour 0 to 12 for human-readable 12-hour time.
5. Compare with the choices, state the final time/option, and explain the key correspondences in 1–2 sentences.

The final answer is placed in <answer></answer>."'

**(3) Abstract Spatial & Logical Puzzles (ASLP)**

**Task: Unfolded Cube**

[Input Image]
prompt = "'This is a cube unfolding problem. Determine which of the options can be folded into the given cube, or what the unfolded pattern looks like. Explain your spatial reasoning.
Question: {question}
The final answer is placed in <answer></answer>."'

**Task: Defuse A Bomb**

[Input Image]
prompt = "'Question: {question} You can first connect the lines to the obstructed area and then go through each option one by one to determine which wire to cut.
The final answer is placed in <answer></answer>."'

**Task: Multi-piece Puzzle**

[Input Image]
prompt = "'Question: {question} You can carefully consider the details of each option before

making your choice.
The final answer is placed in <answer></answer>."'

**Task: Puzzle**

[Input Image]
prompt = "'Given the object above. There is a missing piece in the white area. Which of the five pieces (A, B, C, D, or E) fits perfectly into the missing part of the object? Please examine the immediate surroundings first and work step-by-step: 1. Describe the boundary shape (angles, curves) of the hole. 2. Describe any pattern/stripe/texture crossing the boundary. 3. Note lighting/shading and relative scale. 4. Compare each candidate to steps 1–3 and rule out mismatches. 5. State final choice and a brief justification (3–5 short sentences).
The final answer is placed in <answer></answer>."'

**Task: Trailer Cubes Count**

[Input Image]
prompt = "'Based on the three views, what's the maximum number of cubes that could be present? Steps: 1. For each column (grid position in the top view), determine the maximum possible height consistent with front and side views. 2. Count cubes in each column = column height. 3. Sum across all columns for the total.
The final answer is placed in <answer></answer>."'

**Task: Trailer Cubes Missing**

[Input Image]
prompt = "'Given the three views, what is the minimum number of cubes needed to fill in the structure so that it becomes a solid block with no internal gaps?
Procedure the model must use:

1. Read the top view to list allowed (r,c) column positions.
2. Let $H_{\text{full}}$ be the required cuboid height (the maximum height implied by front/side).
3. To produce a minimal current 3D consistent with views.
4. For each allowed (r,c) column, compute missing $= H_{\text{full}} -$ assigned_height.
5. Sum missing cube values.

The final answer is placed in <answer></answer>."'

**(4) Causal Transformations (CT)**

**Task: Paper Airplane**

[Input Image]
prompt = "'Question: {question}
Please note the differences between the folding positions of the wings, center, and nose of the aircraft in each option, and then choose the appropriate option.
The final answer is placed in <answer></answer>."'

**Task: Gear Rotation**

[Input Image]
prompt = "'Question: {question}
You can answer this question based on the fact that two connected gears rotate in opposite directions, a conveyor belt rotates in the same direction as the gears, and a crossed conveyor belt rotates in the opposite direction as the gears.
The final answer is placed in <answer></answer>."'

**Task: Rolling Dice (Top)**

[Input Image]
prompt = "'Question: {question}
You can list the situation of each side of the dice after each roll, mark the top and bottom, and then after you have reasoned through each step, combine each step with the final result and choose the correct option.
The final answer is placed in <answer></answer>."'

**Task: Rolling Dice (Two)**

[Input Image]
prompt = "'Question: {question}
You can list the situation of each side of the dice after each roll, mark the top and bottom, and then after you have reasoned through each step, combine each step with the final result and choose the correct option.

The final answer is placed in <answer></answer>.”’

**Task: Rolling Dice (Sum)**

[Input Image]
prompt = “‘Question: {question}
You can list the situation of each side of the dice after each roll, mark the top and bottom, and then after you have reasoned through each step, combine each step with the final result and choose the correct option.
The final answer is placed in <answer></answer>.”’

**Level 3: Simulated Visual-CoT Reasoning.** This level evaluates the model's ability to utilize visual information in its reasoning process. Given that current MLLMs are unable to generate their own intermediate visual steps, this setting simulates a Visual-CoT process. The model is provided with the initial problem image along with a sequence of manually annotated intermediate images that act as visual clues. The prompt then directs the model to reason based on this sequence of visuals to arrive at the final answer. This approach is designed to measure the performance improvement gained from visual aids and to understand the potential of a true “think while drawing” capability.

[Input Image] [CoT Image 1] [CoT Image 2] ...
prompt = “‘Based on the question image and the intermediate reasoning image(s) provided, please continue the reasoning to solve the problem.
Question: {question}
The final answer is placed in <answer></answer>.”’

**Evaluation Prompt.** This prompt is used by an evaluator model to judge the correctness of the primary model's response.

[Input Image]
Judge prompt = “‘You are a strict and precise evaluator. Your task is to determine whether the model's final answer is correct based on the ground truth.
Your evaluation must focus exclusively on the answer contained within the <answer></answer>tags, as well as the final answer portion at the end of the model's response. Ignore all reasoning, explanations, or any other text outside of these sections. The correctness of the reasoning process is not part of your evaluation.
Here is the data:
Question: “{question}”
Ground Truth Answer: “{ground truth}”
Model's Full Response: “{model response}”
Based on the ground truth, is the answer inside the <answer>tag correct?
Please respond with only one word: “Correct” or “Incorrect”. ”’

## 8. Detailed Experimental Tables

This section provides a detailed breakdown of model performance across all sub-categories within the **MIRA** benchmark, supplementing the main results presented in Table 1. The following tables correspond to Tables 4-10 as referenced in the main paper.

Table 4. Detailed Results for Euclidean Geometry (Convex Hull, Mirror Pattern) and Physics-Based Reasoning (Mirror Clock) Tasks.

| Model | Convex Hull | | | Mirror Pattern | | | Mirror Clock | | |
|---|---|---|---|---|---|---|---|---|---|
| | D | T | V | D | T | V | D | T | V |
| GPT-5 | 16.7 | 16.7 | 20.0 | 26.7 | 33.3 | 26.7 | 23.3 | 33.3 | 46.7 |
| GPT-5.2 | 46.7 | 26.7 | 38.7 | 40.0 | 40.0 | 30.0 | 23.3 | 23.3 | 66.7 |
| GPT-5-mini | 10.0 | 16.7 | 23.3 | 30.0 | 10.0 | 16.7 | 3.3 | 6.7 | 43.3 |
| GPT-4.1 | 13.3 | 16.7 | 10.0 | 20.0 | 36.7 | 26.7 | 3.3 | 6.7 | 13.3 |
| GPT-4.1-mini | 3.3 | 13.3 | 20.0 | 23.3 | 20.0 | 30.0 | 0.0 | 16.7 | 13.3 |
| GPT-4o | 6.7 | 3.3 | 16.7 | 36.7 | 23.3 | 20.0 | 0.0 | 0.0 | 0.0 |
| GPT-4o-mini | 6.7 | 10.0 | 16.7 | 13.3 | 30.0 | 20.0 | 0.0 | 3.3 | 0.0 |
| o4-mini | 13.8 | 11.5 | 3.3 | 33.3 | 20.0 | 16.7 | 16.7 | 13.3 | 33.3 |
| o3 | 17.9 | 0.0 | 16.7 | 26.7 | 26.7 | 23.3 | 10.0 | 6.7 | 33.3 |
| Claude 4 Opus | 6.7 | 13.3 | 13.3 | 30.0 | 36.7 | 30.0 | 0.0 | 0.0 | 0.0 |
| Claude 4 Sonnet | 16.7 | 10.0 | 13.3 | 26.7 | 23.3 | 20.0 | 6.7 | 3.3 | 6.7 |
| Seed1.5-VL | 13.3 | 13.3 | 16.7 | 16.7 | 16.7 | 26.7 | 0.0 | 0.0 | 16.7 |
| Seed1.6 Vision Pro | 10.0 | 3.3 | 40.0 | 26.7 | 30.0 | 26.7 | 0.0 | 0.0 | 16.7 |
| Gemini 2.5 Flash | 0.0 | 0.0 | 3.3 | 13.3 | 30.0 | 30.0 | 6.7 | 6.7 | 40.0 |
| Gemini 2.5 Pro | 6.7 | 10.0 | 10.0 | 20.0 | 30.0 | 16.7 | 23.3 | 10.0 | 50.0 |
| Gemini 3 Pro | 26.7 | 10.0 | 16.7 | 20.0 | 26.7 | 36.7 | 66.7 | 70.0 | 73.3 |
| Gemini 3 Flash | 46.7 | 20.0 | 46.7 | 10.0 | 23.3 | 10.0 | 76.7 | 70.0 | 90.0 |
| Qwen-VL-Max | 3.3 | 0.0 | 23.3 | 13.3 | 33.3 | 33.3 | 6.7 | 0.0 | 0.0 |
| Qwen3-VL (235B) | 10.0 | 13.3 | 23.3 | 30.0 | 30.0 | 33.3 | 16.7 | 10.0 | 16.7 |
| Qwen2.5-VL (32B) | 0.0 | 0.0 | 10.0 | 13.3 | 3.3 | 13.3 | 0.0 | 0.0 | 3.3 |
| Qwen2.5-VL (72B) | 16.7 | 20.0 | 16.7 | 20.0 | 20.0 | 30.0 | 3.3 | 0.0 | 3.3 |
| GLM 4.5 V (106B) | 16.7 | 13.3 | 20.0 | 30.0 | 33.3 | 26.7 | 0.0 | 0.0 | 0.0 |
| Bagel (7B) | 0.0 | 0.0 | 10.0 | 30.0 | 16.7 | 30.0 | 0.0 | 0.0 | 0.0 |
| Janus-pro (7B) | 0.0 | 16.7 | 0.0 | 3.3 | 13.3 | 23.3 | 0.0 | 0.0 | 0.0 |

Table 5. Detailed Results for Euclidean Geometry (Overlap), Abstract Puzzles (Unfolded Cube), and Physics-Based Reasoning (Billiards) Tasks.

| Model | Overlap | | | Unfolded Cube | | | Billiards | | |
|---|---|---|---|---|---|---|---|---|---|
| | D | T | V | D | T | V | D | T | V |
| GPT-5 | 36.7 | 36.7 | 46.7 | 8.7 | 27.3 | 45.8 | 23.8 | 9.5 | 85.7 |
| GPT-5.2 | 40.0 | 43.3 | 96.7 | 19.2 | 23.1 | 65.4 | 23.8 | 9.5 | 100.0 |
| GPT-5-mini | 20.0 | 36.7 | 80.0 | 0.0 | 13.6 | 50.0 | 9.5 | 9.5 | 61.9 |
| GPT-4.1 | 56.7 | 50.0 | 63.3 | 0.0 | 0.0 | 3.8 | 9.5 | 14.3 | 76.2 |
| GPT-4.1-mini | 3.3 | 13.3 | 20.0 | 23.3 | 20.0 | 30.0 | 0.0 | 16.7 | 13.3 |
| GPT-4o | 53.3 | 36.7 | 30.0 | 0.0 | 0.0 | 0.0 | 4.8 | 14.3 | 57.1 |
| GPT-4o-mini | 13.3 | 40.0 | 16.7 | 0.0 | 0.0 | 0.0 | 19.1 | 9.5 | 38.1 |
| o4-mini | 33.3 | 43.3 | 56.7 | 14.3 | 0.0 | 23.1 | 15.8 | 21.1 | 70.0 |
| o3 | 36.7 | 43.3 | 66.7 | 10.0 | 0.0 | 23.5 | 9.5 | 25.0 | 90.5 |
| Claude 4 Opus | 36.7 | 43.3 | 43.3 | 0.0 | 0.0 | 7.7 | 9.5 | 23.8 | 42.9 |
| Claude 4 Sonnet | 23.3 | 26.7 | 53.3 | 0.0 | 0.0 | 0.0 | 9.5 | 9.5 | 42.9 |
| Seed1.5-VL | 36.7 | 33.3 | 46.7 | 0.0 | 7.7 | 3.8 | 9.5 | 23.8 | 52.4 |
| Seed1.6 Vision Pro | 43.3 | 33.3 | 56.7 | 0.0 | 7.7 | 7.7 | 19.1 | 19.1 | 85.7 |
| Gemini 2.5 Flash | 36.7 | 30.0 | 46.7 | 0.0 | 3.8 | 0.0 | 9.5 | 14.3 | 52.4 |
| Gemini 2.5 Pro | 36.7 | 20.0 | 63.3 | 19.2 | 3.8 | 7.7 | 28.6 | 14.3 | 61.9 |
| Gemini 3 Pro | 53.3 | 40.0 | 66.7 | 38.5 | 26.9 | 23.1 | 33.3 | 42.9 | 95.2 |
| Gemini 3 Flash | 23.3 | 50.0 | 76.7 | 19.2 | 19.2 | 57.7 | 33.3 | 33.3 | 95.2 |
| Qwen-VL-Max | 46.7 | 43.3 | 46.7 | 6.7 | 0.0 | 3.3 | 14.3 | 19.1 | 57.1 |
| Qwen3-VL (235B) | 50.0 | 23.3 | 66.7 | 15.4 | 7.7 | 57.7 | 23.8 | 19.0 | 95.2 |
| Qwen2.5-VL (32B) | 13.3 | 20.0 | 10.0 | 0.0 | 0.0 | 0.0 | 14.3 | 19.1 | 9.5 |
| Qwen2.5-VL (72B) | 40.0 | 43.3 | 40.0 | 0.0 | 0.0 | 0.0 | 4.8 | 9.5 | 76.2 |
| GLM 4.5 V (106B) | 40.0 | 33.3 | 46.7 | 0.0 | 3.8 | 0.0 | 9.5 | 9.5 | 38.1 |
| Bagel (7B) | 10.0 | 13.3 | 16.7 | 0.0 | 0.0 | 0.0 | 19.1 | 0.0 | 19.1 |
| Janus-pro (7B) | 0.0 | 20.0 | 20.0 | 0.0 | 0.0 | 0.0 | 4.8 | 4.8 | 0.0 |

Table 6. Detailed Results for Euclidean Geometry (Localizer), Causal Transformations (Paper Airplane), and Abstract Puzzles (Defuse A Bomb) Tasks.

| Model | Localizer | | | Paper Airplane | | | Defuse A Bomb | | |
|---|---|---|---|---|---|---|---|---|---|
| | D | T | V | D | T | V | D | T | V |
| GPT-5 | 0.0 | 0.0 | 0.0 | 32.0 | 32.0 | 28.0 | 32.0 | 28.0 | 32.0 |
| GPT-5.2 | 0.0 | 0.0 | 0.0 | 20.0 | 12.0 | 8.0 | 12.0 | 20.0 | 16.0 |
| GPT-5-mini | 0.0 | 0.0 | 0.0 | 12.0 | 16.0 | 36.0 | 16.0 | 28.0 | 16.0 |
| GPT-4.1 | 0.0 | 0.0 | 0.0 | 28.0 | 16.0 | 16.0 | 24.0 | 28.0 | 40.0 |
| GPT-4.1-mini | 0.0 | 0.0 | 0.0 | 24.0 | 20.0 | 28.0 | 28.0 | 32.0 | 20.0 |
| GPT-4o | 0.0 | 0.0 | 0.0 | 20.0 | 20.0 | 24.0 | 16.0 | 8.0 | 24.0 |
| GPT-4o-mini | 0.0 | 0.0 | 0.0 | 20.0 | 12.0 | 20.0 | 8.0 | 24.0 | 32.0 |
| o4-mini | 0.0 | 0.0 | 0.0 | 20.0 | 12.0 | 40.0 | 24.0 | 24.0 | 28.0 |
| o3 | 0.0 | 0.0 | 0.0 | 28.0 | 32.0 | 28.0 | 28.0 | 24.0 | 12.0 |
| Claude 4 Opus | 0.0 | 0.0 | 0.0 | 28.0 | 20.0 | 12.0 | 20.0 | 24.0 | 32.0 |
| Claude 4 Sonnet | 0.0 | 0.0 | 0.0 | 16.0 | 12.0 | 12.0 | 28.0 | 36.0 | 28.0 |
| Seed1.5-VL | 0.0 | 0.0 | 0.0 | 4.0 | 24.0 | 16.0 | 32.0 | 40.0 | 8.0 |
| Seed1.6 Vision Pro | 0.0 | 0.0 | 0.0 | 20.0 | 24.0 | 24.0 | 28.0 | 20.0 | 8.0 |
| Gemini 2.5 Flash | 0.0 | 0.0 | 0.0 | 16.0 | 12.0 | 24.0 | 16.0 | 16.0 | 12.0 |
| Gemini 2.5 Pro | 0.0 | 0.0 | 0.0 | 28.0 | 32.0 | 20.0 | 20.0 | 16.7 | 28.0 |
| Gemini 3 Pro | 0.0 | 0.0 | 0.0 | 24.0 | 24.0 | 32.0 | 20.0 | 24.0 | 24.0 |
| Gemini 3 Flash | 0.0 | 0.0 | 0.0 | 28.0 | 40.0 | 36.0 | 24.0 | 28.0 | 28.0 |
| Qwen-VL-Max | 0.0 | 0.0 | 0.0 | 20.0 | 4.0 | 20.0 | 32.0 | 16.0 | 36.0 |
| Qwen3-VL (235B) | 0.0 | 0.0 | 0.0 | 16.0 | 16.0 | 24.0 | 20.0 | 16.0 | 36.0 |
| Qwen2.5-VL (32B) | 0.0 | 0.0 | 0.0 | 12.0 | 12.0 | 16.0 | 0.0 | 0.0 | 4.0 |
| Qwen2.5-VL (72B) | 0.0 | 0.0 | 0.0 | 20.0 | 4.0 | 20.0 | 32.0 | 16.0 | 36.0 |
| GLM 4.5 V (106B) | 0.0 | 0.0 | 0.0 | 4.0 | 12.0 | 12.0 | 30.4 | 20.0 | 28.0 |
| Bagel (7B) | 0.0 | 0.0 | 0.0 | 28.0 | 12.0 | 12.0 | 16.0 | 0.0 | 20.0 |
| Janus-pro (7B) | 0.0 | 0.0 | 0.0 | 4.0 | 8.0 | 12.0 | 16.0 | 0.0 | 8.0 |

Table 7. Detailed Results for Abstract Puzzles (Multi-piece Puzzle), Physics-Based Reasoning (Electric Charge), and Causal Transformations (Rolling Dice: Top) Tasks.

| Model | Multi-piece Puzzle | | | Electric Charge | | | Rolling Dice: Top | | |
|---|---|---|---|---|---|---|---|---|---|
| | D | T | V | D | T | V | D | T | V |
| GPT-5 | 0.0 | 0.0 | 6.7 | 42.7 | 23.8 | 28.6 | 30.8 | 30.8 | 92.3 |
| GPT-5.2 | 3.3 | 6.7 | 10.0 | 61.9 | 42.9 | 66.7 | 26.9 | 34.6 | 100.0 |
| GPT-5-mini | 0.0 | 0.0 | 0.0 | 71.4 | 47.6 | 14.3 | 38.7 | 26.9 | 73.1 |
| GPT-4.1 | 0.0 | 0.0 | 0.0 | 23.8 | 28.6 | 28.6 | 11.5 | 26.9 | 26.9 |
| GPT-4.1-mini | 0.0 | 3.3 | 0.0 | 28.6 | 33.3 | 66.7 | 3.8 | 23.1 | 23.1 |
| GPT-4o | 0.0 | 0.0 | 3.3 | 19.1 | 19.1 | 57.1 | 15.4 | 7.7 | 11.5 |
| GPT-4o-mini | 0.0 | 0.0 | 0.0 | 23.8 | 4.8 | 14.3 | 19.2 | 26.9 | 30.8 |
| o4-mini | 6.9 | 3.6 | 0.0 | 23.8 | 57.1 | 28.6 | 21.1 | 15.0 | 57.7 |
| o3 | 0.0 | 0.0 | 3.6 | 47.6 | 19.1 | 19.1 | 30.8 | 26.9 | 76.9 |
| Claude 4 Opus | 3.3 | 0.0 | 0.0 | 47.6 | 42.9 | 42.9 | 11.5 | 15.4 | 23.1 |
| Claude 4 Sonnet | 3.3 | 3.3 | 0.0 | 42.9 | 42.9 | 33.3 | 15.4 | 34.6 | 30.8 |
| Seed1.5-VL | 0.0 | 0.0 | 0.0 | 52.4 | 61.9 | 61.9 | 23.1 | 38.5 | 19.2 |
| Seed1.6 Vision Pro | 0.0 | 3.3 | 0.0 | 42.9 | 47.6 | 52.4 | 15.4 | 26.9 | 26.9 |
| Gemini 2.5 Flash | 0.0 | 0.0 | 3.3 | 42.9 | 47.6 | 47.6 | 11.5 | 19.2 | 11.5 |
| Gemini 2.5 Pro | 3.3 | 6.7 | 3.3 | 71.4 | 57.1 | 66.7 | 7.7 | 23.1 | 15.4 |
| Gemini 3 Pro | 0.0 | 0.0 | 0.0 | 71.4 | 57.1 | 71.4 | 34.6 | 53.8 | 23.1 |
| Gemini 3 Flash | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.5 | 19.2 | 15.4 | 38.5 |
| Qwen-VL-Max | 3.3 | 0.0 | 0.0 | 52.4 | 47.6 | 38.1 | 19.2 | 19.2 | 38.5 |
| Qwen3-VL (235B) | 0.0 | 0.0 | 3.3 | 66.7 | 61.9 | 61.9 | 19.2 | 23.1 | 42.3 |
| Qwen2.5-VL (32B) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7.7 | 11.5 | 7.7 |
| Qwen2.5-VL (72B) | 0.0 | 0.0 | 3.3 | 57.1 | 47.6 | 47.6 | 23.1 | 11.5 | 23.1 |
| GLM 4.5 V (106B) | 0.0 | 3.3 | 0.0 | 42.9 | 52.4 | 33.3 | 19.2 | 26.9 | 61.5 |
| Bagel (7B) | 0.0 | 0.0 | 0.0 | 4.8 | 0.0 | 0.0 | 15.4 | 3.8 | 19.2 |
| Janus-pro (7B) | 0.0 | 0.0 | 0.0 | 0.0 | 14.3 | 0.0 | 11.5 | 3.8 | 11.5 |

Table 8. Detailed Results for Causal Transformations Tasks (Rolling Dice: Sum, Rolling Dice: Two, Gear Rotation).

| Model | Rolling Dice: Sum | | | Rolling Dice: Two | | | Gear Rotation | | |
|---|---|---|---|---|---|---|---|---|---|
| | D | T | V | D | T | V | D | T | V |
| GPT-5 | 11.5 | 3.8 | 8.0 | 0.0 | 0.0 | 0.0 | 15.0 | 30.0 | 15.0 |
| GPT-5.2 | 7.7 | 11.5 | 15.4 | 0.0 | 7.7 | 73.1 | 25.0 | 15.0 | 30.0 |
| GPT-5-mini | 15.4 | 7.7 | 3.8 | 0.0 | 0.0 | 0.0 | 20.0 | 15.0 | 10.0 |
| GPT-4.1 | 11.5 | 11.5 | 3.8 | 0.0 | 0.0 | 0.0 | 15.0 | 35.0 | 30.0 |
| GPT-4.1-mini | 3.8 | 3.8 | 7.7 | 0.0 | 0.0 | 0.0 | 20.0 | 30.0 | 15.0 |
| GPT-4o | 0.0 | 7.7 | 0.0 | 0.0 | 0.0 | 0.0 | 35.0 | 25.0 | 10.0 |
| GPT-4o-mini | 3.8 | 7.7 | 0.0 | 0.0 | 0.0 | 0.0 | 35.0 | 40.0 | 25.0 |
| o4-mini | 11.8 | 5.0 | 4.2 | 0.0 | 0.0 | 0.0 | 30.0 | 40.0 | 20.0 |
| o3 | 11.5 | 12.0 | 7.7 | 0.0 | 0.0 | 0.0 | 30.0 | 30.0 | 25.0 |
| Claude 4 Opus | 3.8 | 7.7 | 15.4 | 0.0 | 0.0 | 0.0 | 20.0 | 15.0 | 10.0 |
| Claude 4 Sonnet | 11.5 | 3.8 | 0.0 | 0.0 | 0.0 | 0.0 | 20.0 | 25.0 | 5.0 |
| Seed1.5-VL | 7.7 | 7.7 | 7.7 | 0.0 | 0.0 | 0.0 | 35.0 | 20.0 | 20.0 |
| Seed1.6 Vision Pro | 3.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 45.0 | 0.0 | 5.0 |
| Gemini 2.5 Flash | 7.7 | 3.8 | 0.0 | 0.0 | 0.0 | 0.0 | 35.0 | 25.0 | 35.0 |
| Gemini 2.5 Pro | 15.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 35.0 | 30.0 | 15.0 |
| Gemini 3 Pro | 3.8 | 11.5 | 0.0 | 0.0 | 0.0 | 3.8 | 75.0 | 70.0 | 80.0 |
| Gemini 3 Flash | 11.5 | 3.8 | 11.5 | 7.7 | 0.0 | 61.5 | 15.0 | 25.0 | 30.0 |
| Qwen-VL-Max | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7.7 | 30.0 | 15.0 | 35.0 |
| Qwen3-VL (235B) | 0.0 | 3.8 | 3.8 | 0.0 | 0.0 | 26.9 | 15.0 | 30.0 | 10.0 |
| Qwen2.5-VL (32B) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 30.0 | 0.0 |
| Qwen2.5-VL (72B) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 35.0 | 5.0 |
| GLM 4.5 V (106B) | 15.4 | 3.8 | 11.5 | 7.7 | 0.0 | 34.6 | 20.0 | 25.0 | 10.0 |
| Bagel (7B) | 3.8 | 0.0 | 0.0 | 3.3 | 3.3 | 3.3 | 15.4 | 23.1 | 11.5 |
| Janus-pro (7B) | 3.8 | 7.7 | 0.0 | 0.0 | 0.0 | 0.0 | 25.0 | 10.0 | 5.0 |

Table 9. Detailed Results for Euclidean Geometry (Cubes Count, Cubes Missing) and Abstract Puzzles (Puzzle) Tasks.

| Model | Cubes Count | | | Cubes Missing | | | Puzzle | | |
|---|---|---|---|---|---|---|---|---|---|
| | D | T | V | D | T | V | D | T | V |
| GPT-5 | 3.3 | 0.0 | 0.0 | 3.3 | 0.0 | 0.0 | 3.8 | 34.6 | 30.8 |
| GPT-5.2 | 0.0 | 0.0 | 0.0 | 0.0 | 3.3 | 0.0 | 23.1 | 26.9 | 30.8 |
| GPT-5-mini | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 19.2 | 19.2 | 23.1 |
| GPT-4.1 | 0.0 | 0.0 | 0.0 | 6.7 | 3.3 | 0.0 | 11.5 | 15.4 | 15.4 |
| GPT-4.1-mini | 0.0 | 0.0 | 0.0 | 3.3 | 6.7 | 0.0 | 23.1 | 15.4 | 15.4 |
| GPT-4o | 3.3 | 0.0 | 3.3 | 3.3 | 3.3 | 0.0 | 11.5 | 11.5 | 30.8 |
| GPT-4o-mini | 3.3 | 0.0 | 6.7 | 6.7 | 3.3 | 6.7 | 34.6 | 15.4 | 23.1 |
| o4-mini | 3.6 | 4.0 | 7.1 | 0.0 | 0.0 | 0.0 | 30.8 | 23.1 | 19.2 |
| o3 | 3.3 | 6.7 | 0.0 | 6.7 | 3.3 | 3.3 | 26.9 | 23.1 | 34.6 |
| Claude 4 Opus | 0.0 | 0.0 | 0.0 | 3.3 | 0.0 | 3.3 | 19.2 | 23.1 | 23.1 |
| Claude 4 Sonnet | 3.3 | 0.0 | 3.3 | 3.3 | 0.0 | 0.0 | 30.8 | 26.9 | 23.1 |
| Seed1.5-VL | 0.0 | 0.0 | 3.3 | 0.0 | 0.0 | 3.3 | 3.8 | 7.7 | 7.7 |
| Seed1.6 Vision Pro | 0.0 | 0.0 | 3.3 | 0.0 | 0.0 | 3.3 | 11.5 | 0.0 | 3.8 |
| Gemini 2.5 Flash | 3.3 | 0.0 | 0.0 | 3.3 | 10.0 | 13.3 | 19.2 | 15.4 | 19.2 |
| Gemini 2.5 Pro | 0.0 | 6.7 | 0.0 | 0.0 | 0.0 | 0.0 | 11.5 | 11.5 | 19.2 |
| Gemini 3 Pro | 3.3 | 0.0 | 3.3 | 0.0 | 0.0 | 3.3 | 34.6 | 30.8 | 34.6 |
| Gemini 3 Flash | 3.3 | 0.0 | 3.3 | 6.7 | 3.3 | 6.7 | 23.1 | 23.1 | 38.5 |
| Qwen-VL-Max | 6.7 | 0.0 | 3.3 | 0.0 | 0.0 | 0.0 | 30.8 | 34.6 | 23.1 |
| Qwen3-VL (235B) | 0.0 | 0.0 | 3.3 | 3.3 | 3.3 | 10.0 | 23.1 | 34.6 | 26.9 |
| Qwen2.5-VL (32B) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7.7 | 23.1 | 19.2 |
| Qwen2.5-VL (72B) | 6.7 | 0.0 | 0.0 | 3.3 | 0.0 | 0.0 | 30.8 | 23.1 | 23.1 |
| GLM 4.5 V (106B) | 0.0 | 0.0 | 0.0 | 3.3 | 3.3 | 3.3 | 19.2 | 15.4 | 23.1 |
| Bagel (7B) | 0.0 | 0.0 | 3.3 | 3.3 | 3.3 | 3.3 | 15.4 | 23.1 | 11.5 |
| Janus-pro (7B) | 0.0 | 0.0 | 0.0 | 10.0 | 0.0 | 6.7 | 0.0 | 34.6 | 26.9 |

Table 10. Detailed Results for Abstract Puzzles Tasks (Trailer Cubes Count, Trailer Cubes Missing).

| Model | Trailer Cubes Count | | | Trailer Cubes Missing | | |
|---|---|---|---|---|---|---|
| | D | T | V | D | T | V |
| GPT-5 | 16.0 | 4.0 | 4.0 | 4.0 | 0.0 | 0.0 |
| GPT-5.2 | 52.0 | 44.0 | 56.0 | 0.0 | 0.0 | 0.0 |
| GPT-5-mini | 4.0 | 4.0 | 8.0 | 4.0 | 0.0 | 4.0 |
| GPT-4.1 | 0.0 | 0.0 | 4.0 | 4.0 | 4.0 | 0.0 |
| GPT-4.1-mini | 0.0 | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 |
| GPT-4o | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GPT-4o-mini | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 | 0.0 |
| o4-mini | 11.8 | 17.7 | 0.0 | 0.0 | 0.0 | 0.0 |
| o3 | 4.0 | 0.0 | 0.0 | 0.0 | 4.0 | 4.0 |
| Claude 4 Opus | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 | 0.0 |
| Claude 4 Sonnet | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Seed1.5-VL | 16.0 | 12.0 | 0.0 | 0.0 | 0.0 | 4.0 |
| Seed1.6 Vision Pro | 4.0 | 12.0 | 8.0 | 8.0 | 8.0 | 0.0 |
| Gemini 2.5 Flash | 0.0 | 0.0 | 4.0 | 4.0 | 0.0 | 4.0 |
| Gemini 2.5 Pro | 8.0 | 0.0 | 0.0 | 4.0 | 4.0 | 0.0 |
| Gemini 3 Pro | 56.0 | 56.0 | 24.0 | 0.0 | 4.0 | 4.0 |
| Gemini 3 Flash | 40.0 | 44.0 | 52.0 | 4.0 | 4.0 | 0.0 |
| Qwen-VL-Max | 4.0 | 0.0 | 4.0 | 4.0 | 4.0 | 4.0 |
| Qwen3-VL (235B) | 8.0 | 0.0 | 0.0 | 4.0 | 0.0 | 0.0 |
| Qwen2.5-VL (32B) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4.0 |
| Qwen2.5-VL (72B) | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 | 0.0 |
| GLM 4.5 V (106B) | 4.0 | 4.0 | 4.0 | 0.0 | 0.0 | 8.0 |
| Bagel (7B) | 0.0 | 16.0 | 0.0 | 0.0 | 4.0 | 4.0 |
| Janus-pro (7B) | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 | 4.0 |

## 9. Dataset Showcase

The **MIRA** benchmark is composed of 546 multimodal problems spanning 20 distinct task types. These tasks are curated to be challenging and require intermediate visual reasoning, a process analogous to how humans "draw to think" to solve complex problems. The tasks fall into four challenging domains: Euclidean Geometry (EG), Physics-Based Reasoning (PBR), Abstract Spatial & Logical Puzzles (ASLP), and Causal Transformations (CT).

To supplement the overview provided in Figure 1 and offer a more intuitive understanding of the dataset, we showcase several representative examples for each category below (Figure 6- 15).

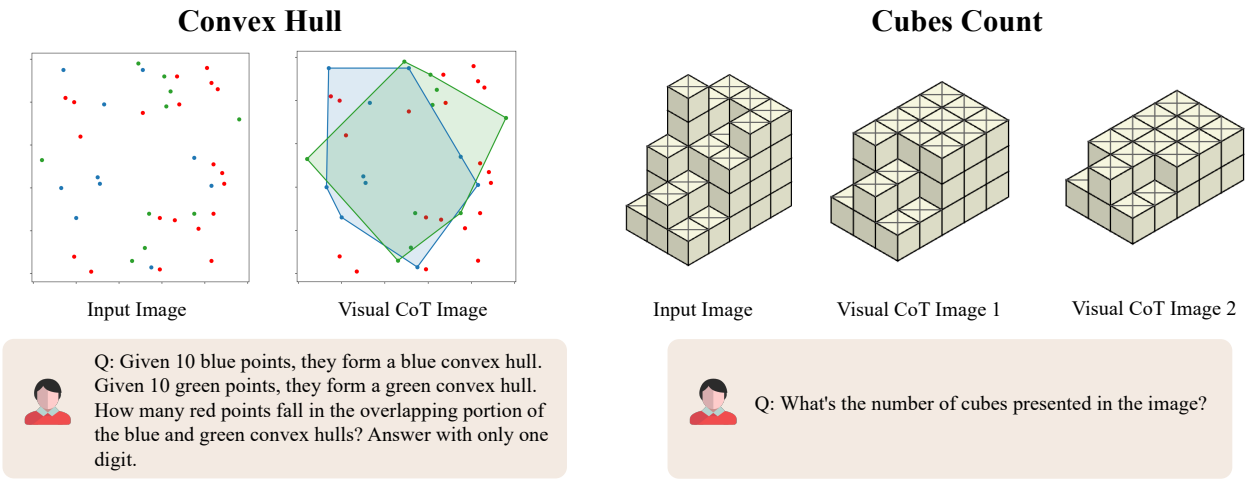



Figure 6. Illustrative cases for Convex Hull task (left) and Cubes Count task (right).

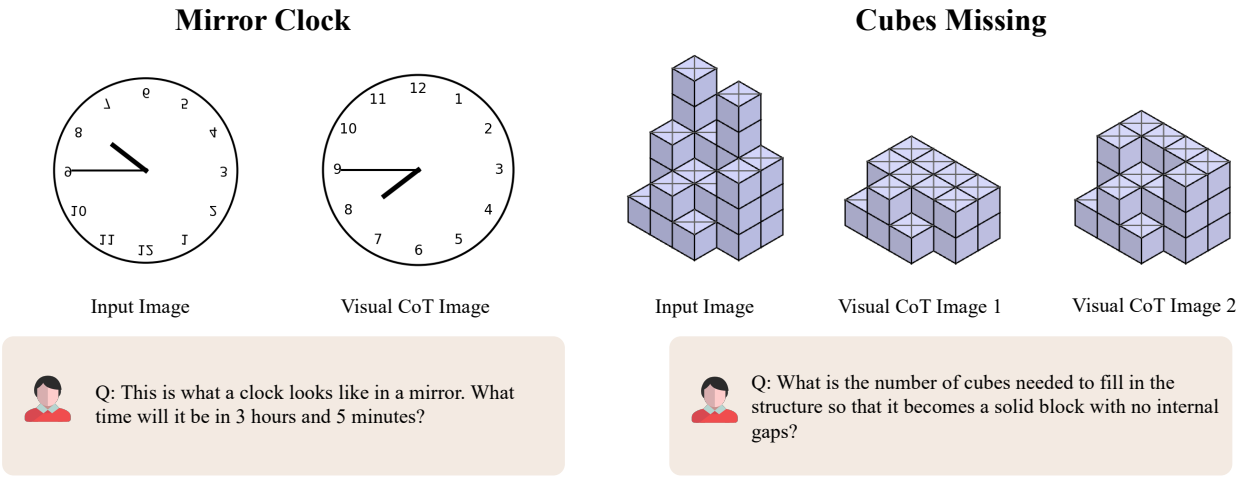



Figure 7. Illustrative cases for Mirror Clock task (left) and Cubes Missing task (right).

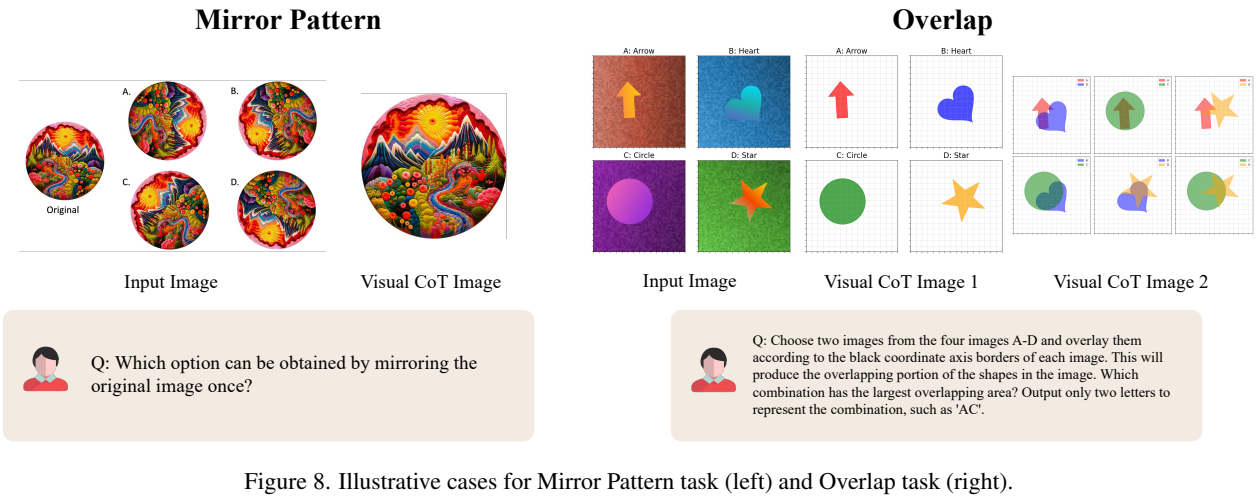
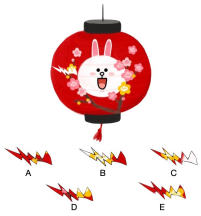



Figure 8. Illustrative cases for Mirror Pattern task (left) and Overlap task (right).

## Puzzle




Input Image

Visual CoT Image

Q: Given the object above. There is a missing piece in the white area. Which of the five pieces (A, B, C, D, or E) fits perfectly into the missing part of the object?

## Trailer Cubes Count


Input Image | Visual CoT Image 1 | Visual CoT Image 2

Q: Based on the three views, what's the maximum number of cubes that could be present?

Figure 9. Illustrative cases for Puzzle task (left) and Trailer Cubes Count task (right).

## Billiards

Input Image | Visual CoT Image

Q: In the image, a billiards table is shown with its pockets numbered 1 – 6 in red. If the blue ball rolls in the direction indicated by the green arrow—without spin and with unlimited momentum—and is guaranteed to drop into a pocket, which pocket will it enter? Please answer with the single digit 1 – 6.

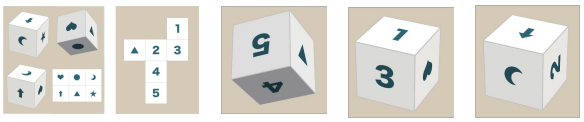
## Trailer Cubes Missing


Input Image | Visual CoT Image 1 | Visual CoT Image 2

Q: Given the three views, what is the minimum number of cubes needed to fill in the structure so that it becomes a solid block with no internal gaps?

Figure 10. Illustrative cases for Puzzle task (left) and Trailer Cubes Count task (right).

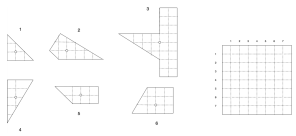
## Unfolded Cube



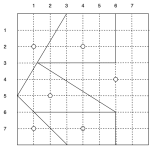
Input Image | CoT Image 1 | CoT Image 2 | CoT Image 3

Q: Given three different perspectives of a cube (top left, top right, and bottom left in the left image) and the patterns on its six faces (bottom right in the left image), please fill in the corresponding patterns in its unfolded view (right image). Note that only the outer surface of the cube has patterns; the inner surface does not. That is, the unfolded view on the right represents the outer surface of the cube. The six patterns are represented by 'heart, 'circle,' 'moon,' 'arrow,' 'triangle,' and 'star,' respectively. The output format is '(number, shape),' with different correspondences separated by semicolons.

## Localizer



Input Image

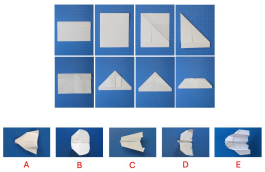


Visual CoT Image

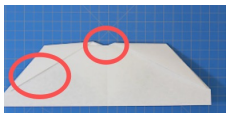
Q: Tile the square on the right using the solid-outlined puzzle pieces on the left. Use all pieces; the final tiling must be exact no leftovers, no gaps, and no overlaps. Each piece has a circle. After completing the tiling, return the coordinates of each circle in numerical order, using the format: [piece ID (e.g., 1), (x, y)]; separate entries with semicolons.

Figure 11. Illustrative cases for Unfolded Cube task (left) and Localizer task (right).

## Paper Airplane



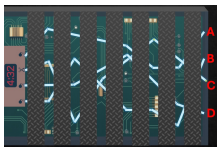

Input Image



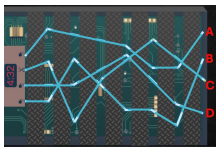
Visual CoT Image

Q: Based on the partial folding steps shown at the top of the image (ordered from left to right, top to bottom), can you infer which of the paper airplanes below is the final result?

## Defuse A Bomb




Input Image



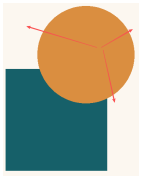

Visual CoT Image

Q: Bomb in a Box — which wire will you cut to defuse it? You can't open the box or see beneath the metal slats, and the wires only change direction where they're visible.

Figure 12. Illustrative cases for Paper Airplane task (left) and Defuse A Bomb task (right).

## Multi-piece Puzzle

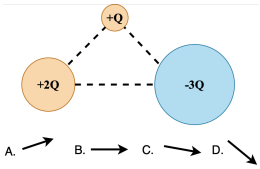
Input Image



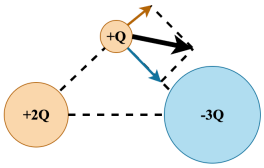
Visual CoT Image

Q: Which three of the nine pieces on the right can be combined to exactly match the shape on the left? Please list the three piece numbers (comma-separated, ascending order).

## Electric Charge




Input Image

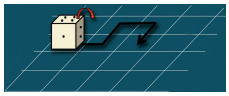



Visual CoT Image

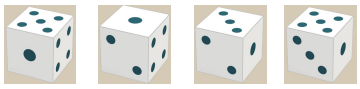
Q: Analyze the forces acting on the topmost positive charge (+Q) and indicate the direction of the net force on it. Note: the size of each sphere does not represent the charge magnitude; use the label inside each sphere to determine its charge.

Figure 13. Illustrative cases for Multi-piece Puzzle task (left) and Electric Charge task (right).

## Rolling Dice: Top



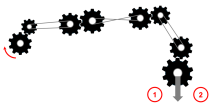
Input Image



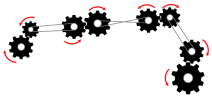
CoT Image 1 CoT Image 2 CoT Image 3 CoT Image 4

Q: If the dice is rolled on the showed path, what will be the number on the top?

## Gear Rotation




Input Image



Visual CoT Image

Q1: If the first gear in the image rotates slightly in the direction of the arrow, will the arrow on the final gear point to 1 or 2? Please answer with a number.

Q2: If the first gear in the image rotates slightly in the direction opposite to the arrow, will the arrow on the final gear point to 1 or 2? Please answer with a number.

Figure 14. Illustrative cases for Rolling Dice: Top task (left) and Gear Rotation task (right).

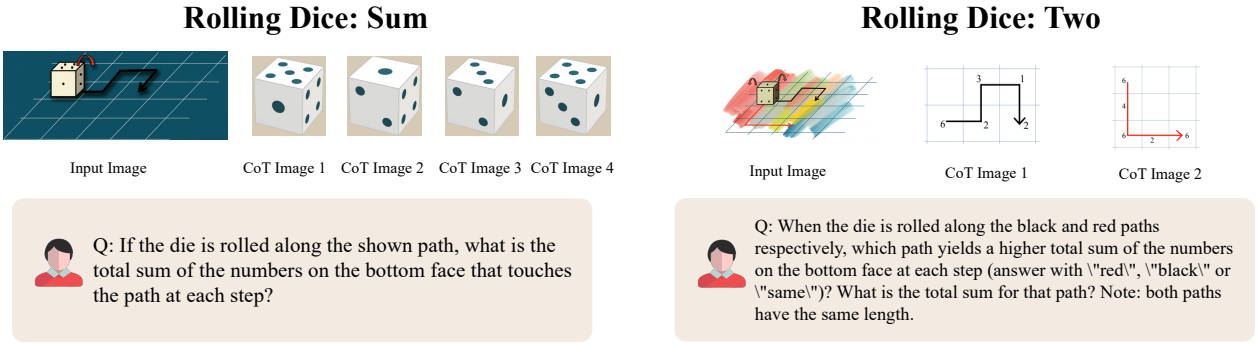



Figure 15. Illustrative cases for Rolling Dice: Sum task (left) and Rolling Dice: Two task (right).