Think Visually, Reason Textually: Vision-Language Synergy in Abstract Reasoning

Supplementary Material

A. Prompts

Prompt for Text-only Reasoning in ARC-AGI:
I will provide you with several input and output matrices. You need to find the matrix-changing rule from it and apply it to the new input. Put the output matrix within \\boxed{}.

Example Input 1: [[0, 1, 0, ...], [0, 1, 1, ...], ...]
Example Output 1: [[1, 1, 0, ...], [1, 1, 1, ...], ...]
Example Input 2: [[0, 2, 4, ...], [0, 2, 2, ...], ...]
Example Output 2: [[2, 2, 4, ...], [2, 2, 2, ...], ...]
...
Example Input n: [[0, 4, 0, ...], [0, 6, 1, ...], ...]
Example Output n: [[4, 4, 0, ...], [6, 6, 1, ...], ...]
New Input: [[0, 2, 0, ...], [0, 5, 3, ...], ...]

Prompt for Vision-centric Rule Summarization:
I will now provide you with several input and output images about 2D grids. You need to summarize the grid-changing rule from it. Output the rule you learned within \\boxed{}.

Example Input 1: <Input Image 1>
Example Output 1: <Output Image 1>
Example Input 2: <Input Image 2>
Example Output 2: <Output Image 2>
...
Example Input n: <Input Image n>
Example Output n: <Output Image n>

Prompt for Text-centric Rule Application:
I will provide you with several input and output matrices. You need to find the matrix-changing rule from it and apply it to the new input. Put the output matrix within \\boxed{}.

Here is a possible rule for your reference. *Rule: The rule involves removing the colored cross ...* Note that the rule is described in color and each color represents a value in the matrix: [0:black; 1:blue; 2:red; 3:green; 4:yellow; 5:grey; 6:pink; 7:orange; 8:light blue; 9:brown]. You need to first check the correctness of the rule based on the examples. If the rule is correct, apply it to the new input. Otherwise, summarize a new rule and apply it to the new input.

Example Input 1: [[0, 1, 0, ...], [0, 1, 1, ...], ...]
Example Output 1: [[1, 1, 0, ...], [1, 1, 1, ...], ...]
Example Input 2: [[0, 2, 4, ...], [0, 2, 2, ...], ...]
Example Output 2: [[2, 2, 4, ...], [2, 2, 2, ...], ...]
...
Example Input n: [[0, 4, 0, ...], [0, 6, 1, ...], ...]
Example Output n: [[4, 4, 0, ...], [6, 6, 1, ...], ...]
New Input: [[0, 2, 0, ...], [0, 5, 3, ...], ...]

Prompt for Vision-centric Consistency Verification:
I will now provide you with several input and output example images, which follows a specific changing rule. Then, I will give you another input and output pair, determine whether the new pair also follows the same changing rule. Add your final judgment at the end of your replay: \\boxed{True} or \\boxed{False}.

Example Input 1: <Input Image 1>
Example Output 1: <Output Image 1>
Example Input 2: <Input Image 2>
Example Output 2: <Output Image 2>
...
Example Input n: <Input Image n>
Example Output n: <Output Image n>
New Input: <New Input Image>
New Output: <Output Image Pred>

B. Matrix-to-Image Visualization

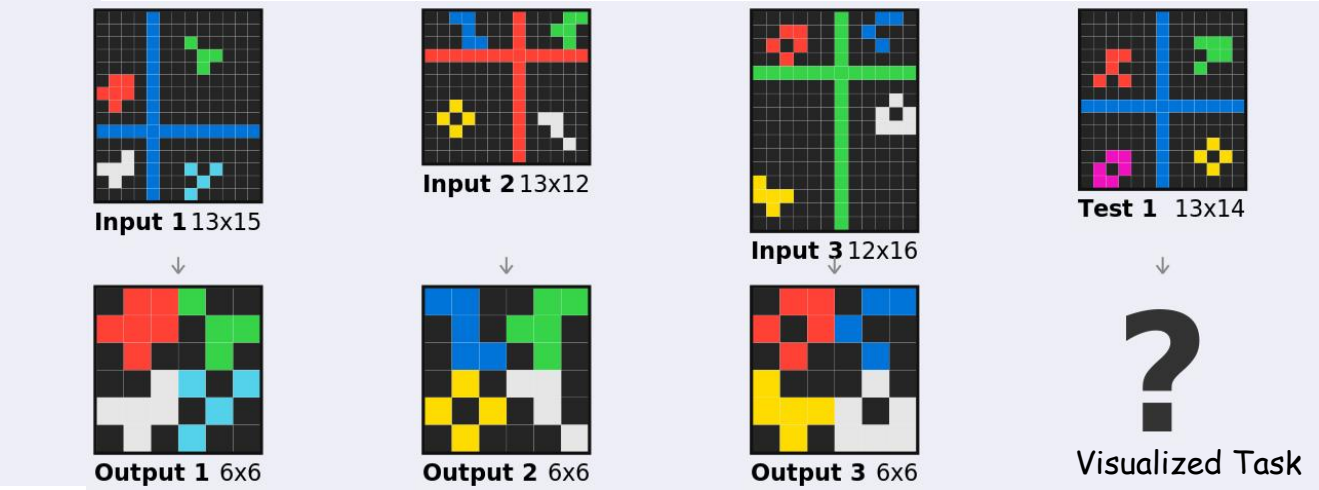
We visualize the input-output matrices into color-coded 2D grids to provide 2D spatial information and global view. The detailed visualization process is listed as follows.

First, each value is mapped to a distinct color. The color map is: [0: black; 1: blue; 2: red; 3: green; 4: yellow; 5: grey; 6: pink; 7: orange; 8: light blue; 9: brown]

Furthermore, between the two elements (small colored squares), we add white dividing lines to more clearly indicate the specific number and the structure of the elements contained within a block.

C. Qualitative Examples

We conduct an in-depth analysis of the specific outputs of different models (GPT-4o, Gemini-2.5-Pro-thinking-8192, o4-mini) when employing visual thinking versus textual thinking in the ARC-AGI task. Visual thinking demonstrates numerous unique advantages, such as the integration of 2D structural information(Fig. 6), a global perspective(Fig. 5, Fig. 7), and long-range perception capabilities(Fig. 8). Therefore, using visual information during the global rule summarization phase enables more effective extraction of the underlying rules.



To change the input matrix, remove all-zero rows and columns. Identify the non-zero submatrix and remove any surrounding zero rows and columns. Essentially, trim the matrix to its smallest size that contains all non-zero elements, preserving the order and positions of these elements relative to each other.

Text-Only Reasoning

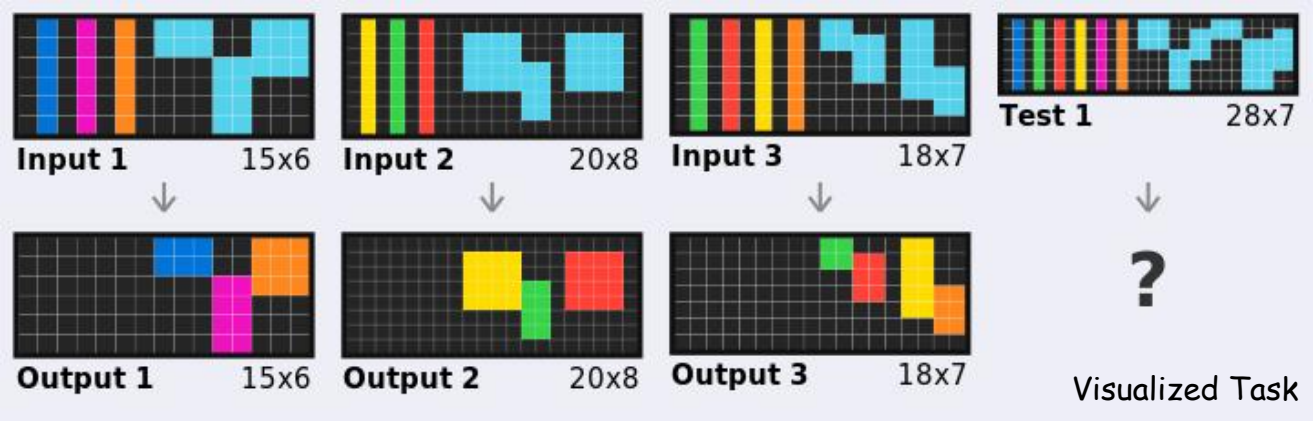
The rule involves removing the colored cross in the middle and shifting the remaining colored shapes towards the center to fill the space evenly. Each quadrant's shape moves towards the center, maintaining its orientation.

Vision-Centric Rule Summarization

Visual thinking is holistic, enabling it to quickly identify the most critical feature (the colored cross) and subsequently summarize the correct rule: assembling the shapes in each quadrant toward the center. In contrast, textual thinking is independent and lacks such a holistic perspective and fails to capture the prominent feature of the colored cross, thereby erroneously interpreting the task as compressing the matrix to include all non-zero elements.

Main Difference

Figure 5. Visual reasoning possesses a global perspective, enabling it to better capture the most critical feature (the colored cross) in the entire image and subsequently summarize the correct rule the underlying rules. Base model is both gpt-4o.



Here's the idea in a nutshell:
1. In the input, the 8's mark "target" slots into which the other non-zero values will move. All original non-8 entries and all the 8's are eventually zeroed out.
2. Work row by row. In each row a) read off its non-zero entries (ignoring any 8's), keeping them in left-to-right order; b) find all contiguous runs of 8's in that row (again, left to right); c) for the k_th run of 8's, replace those 8's by the k_th non-zero entry from (a)
3. Everything else becomes 0.

Text-Only Reasoning

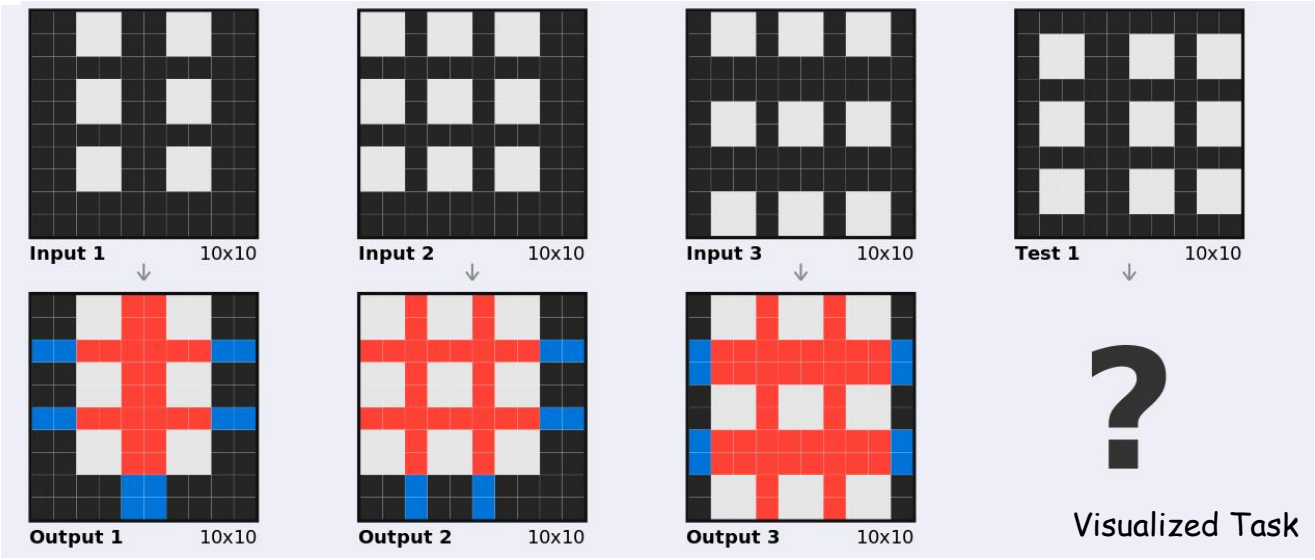
Take the ordered list of colored bars on the left, delete the bars, and then repaint the cyan "Tetris" blocks on the right by cycling through exactly those same colors, in the same left-to-right order. In other words, remove the yellow/green/red (etc.) stripes, and color the first cyan block (or pixel) you encounter with the first stripe-color, the next cyan block with the second stripe-color, and so on, leaving the background black and discarding the original bars.

Vision-Centric Rule Summarization

Textual reasoning tends to process the input matrix *"row by row"*, thus failing to summarize the correct rules. It incorrectly replace every element in a row based on the elements on the left, whereas in reality, the replacement of elements in each row is still related to the presence of blocks in the previous rows. In contrast, visual reasoning can flexibly process the matrix *"block by block"* and correctly summarize the variation rule: coloring is performed block by block from left to right.

Main Difference

Figure 6. Visual reasoning possesses 2D information and can flexibly summarize rules in a "block-by-block" manner, whereas text reasoning tends to adopt a "row-by-row" processing approach, thus failing to derive the correct rules. Base model is both o4-mini.



If the input value is 5, it remains 5.
If the input value is 0, its new value is determined by its 8 neighbors (including diagonals):
If it has 5s as neighbors both vertically (up/down) AND horizontally (left/right), it becomes 2.
If it has 5s as neighbors either vertically OR horizontally (but not both), it becomes 1.
Otherwise, it remains 0.

Text-Only Reasoning

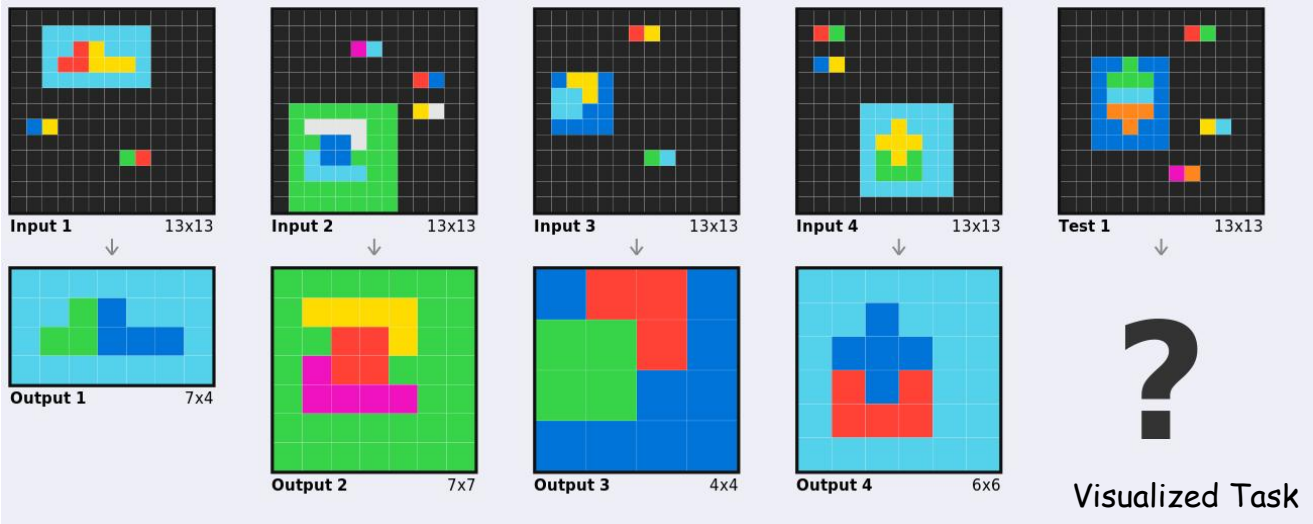
The rule is to recolor the black paths in the input image based on their position relative to the grey squares.
1. **Identify Internal Paths:** Any black path segment that is located between two grey squares is colored red. This applies to both horizontal paths that separate two rows of squares and vertical paths that separate two columns of squares.
2. **Identify External Paths:** Any black path segment that is located between a grey square on one side and the outer black background on the other side is colored blue. These are the paths on the periphery of the grid of squares.
3. **No Change:** The grey squares themselves and the black background remain unchanged.

Vision-Centric Rule Summarization

Visual thinking possesses a superior global perspective, enabling the identification of the core feature (grey squares) in the figure and subsequent processing based on Internal Paths and External Paths. In contrast, textual thinking focuses more on local information, thus conducting processing through the 8 neighbors corresponding to a specific element.

Main Difference

Figure 7. Visual thinking tends to adopt a global perspective and thus processes information based on internal and external paths; in contrast, textual thinking focuses more on local information and processes individual elements with reference to their 8-neighbor context. Base model is both Gemini-2.5-Pro-thinking-8192.



First, find the largest contiguous non-zero area in the input grid and crop to its bounding box. For each cell in this new grid, its new number is determined by its four cardinal neighbors (up, down, left, right). The new number is the count of unique values among these four neighbors that are different from the cell's own original value. If a cell is on the border, only its existing neighbors are counted.

Text-Only Reasoning

The image transformation rule is a color replacement process guided by key-value pairs. In each input image, there is a central object composed of colored shapes within a larger colored frame, and several small two-color pairs scattered on the black background. These pairs act as a color legend. For each pair, the color of the right square indicates an original color to be found within the central object's inner shapes, and the color of the left square is the new color it should be changed to. The transformation consists of applying these color replacements to the shapes inside the frame, while the frame's color remains unchanged. The final output displays only the recolored central object, with the background and the color-key pairs removed.

Vision-Centric Rule Summarization

Textual thinking represents each element individually, thus lacking "long-range correlation" and potentially overlooking some detailed elements. In the context of this example, textual thinking fails to accurately grasp the value of small two-number pairs; in contrast, visual thinking can fully comprehend the function of two-color pairs and summarize the correct rules.

Main Difference

Figure 8. Visual thinking possesses superior long-range correlation capabilities and can better capture detailed features (the 2-color pairs and the re-color strategy). Base model is both Gemini-2.5-Pro-thinking-8192.