TiViBench: Benchmarking Think-in-Video Reasoning for Video Generation

Supplementary Material

A. More Details of Evaluation Dimension

A.1. Motivation for Each Scenario

To comprehensively evaluate visual reasoning in video generative models, we design 24 diverse task scenarios across four key dimensions. Each scenario is carefully crafted to challenge specific aspects of visual reasoning, ensuring a systematic assessment of models' ability to perform beyond general video generation. Below, we outline the motivation for each scenario:

Structural Reasoning & Search. Structural reasoning tasks assess models' ability to understand and navigate complex spatial structures, sequences, and rules, which are critical for reasoning in dynamic environments.

- *Graph Traversal*: Tests the model's capability to explore structured graphs and identify valid traversal paths, simulating real-world spatial reasoning.
- *Maze Solving*: Challenges models to navigate through constrained environments, requiring spatial planning and decision-making.
- *Sorting Numbers*: Evaluates logical ordering of visual elements, emphasizing reasoning over numerical structures in dynamic contexts.
- *Temporal Ordering*: Assesses the model's ability to infer sequential relationships between events or frames.
- *Rule Extrapolation*: Tests the model's understanding of abstract rules and its ability to generalize them to new scenarios.
- *Game Move*: Simulates decision-making in strategic games, requiring models to predict valid moves based on spatial and logical reasoning.

Spatial & Visual Pattern Reasoning. These scenarios focus on recognizing patterns, relationships, and visual consistencies, which are foundational to reasoning in visual contexts.

- *Shape Fitting*: Challenges models to match shapes into predefined spaces, testing spatial alignment and pattern recognition.
- *Connecting Colors*: Evaluates the ability to identify and connect visually related elements based on color patterns.
- *Pattern Recognition*: Assesses model's capacity to detect recurring patterns and infer underlying rules.
- *Odd-one-out*: Tests model's ability to identify anomalies in visual sets, requiring attention to detail and comparative reasoning.
- *Counting Objects*: Focuses on numerical reasoning by evaluating the model's ability to count and quantify visual elements.
- *Visual Analogy*: Assesses abstract reasoning by requiring models to identify analogical relationships between visual objects.

Symbolic & Logical Reasoning. Symbolic reasoning tasks test the ability to understand abstract symbols, logical rules, and numerical relationships.

- *Sudoku Completion*: Challenges models to complete structured puzzles based on logical constraints, testing symbolic reasoning.
- *Symbolic Reasoning*: Evaluates the model's ability to infer relationships between abstract symbols and make logical deductions.
- *Arithmetic*: Tests numerical reasoning by requiring models to solve basic arithmetic problems presented visually.
- *Visual Deduction*: Assesses the ability to infer logical conclusions from visual cues, such as completing partially visible objects.
- *Transitive Reasoning*: Challenges models to infer indirect relationships between elements based on transitive logic.
- *Game Rule*: Evaluates understanding of abstract rules and their application in visual environments.

Action Planning & Task Execution. These tasks simulate real-world scenarios requiring multi-step planning, execution, and adaptability in dynamic environments.

- *Tool Use*: Assesses models' ability to infer the correct use of tools based on visual cues and task requirements.
- *Robot Navigation*: Challenges models to plan and execute navigation in complex spatial environments, simulating robotic reasoning.
- *Goal-directed Planning*: Tests multi-step planning towards achieving specific goals in dynamic settings.
- *Multi-step Manipulation*: Evaluates the ability to coordinate and execute sequential actions to manipulate objects.
- *Instruction Following*: Assesses models' capacity to interpret visual instructions and execute tasks accordingly.
- *Game Strategy*: Challenges strategic reasoning by requiring models to plan and execute moves in visually dynamic games.

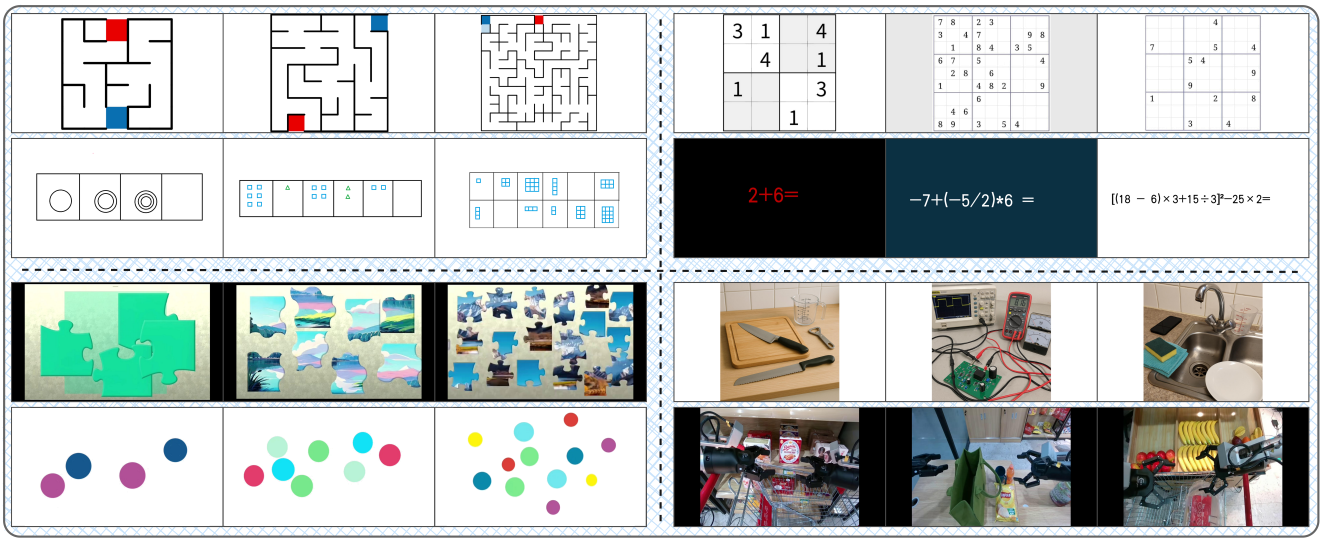


Figure 8. Data demonstration across easy, medium, and hard. (***Top Left***) Structural Reasoning & Search. (***Top Right***) Symbolic & Logical Reasoning. (***Bottom Left***) Spatial & Visual Pattern Reasoning. (***Bottom Right***) Action Planning & Task Execution.

A.2. Data Demonstration

Here, we present examples of our **TiViBench** in Figure 8 to provide a more vivid illustration of the three difficulty levels: easy, medium, and hard.

B. More Details of Prompt Suite

B.1. Prompt Generation

We adopt Gemini-2.5-Pro [9] as a powerful assistant to generate an initial version of the inference prompt for our **TiViBench**. Here we further detail the task prompt for each dimension:

Structural Reasoning & Search

```
"""You are a senior researcher in computer vision
     . You are tasked with generating detailed
     prompts for Image-to-Video (I2V) data
     samples that evaluate Structural Reasoning &
      Search abilities.
You are given two images: {initial_image} shows
     the initial state, and {target_image} shows
     the target state. The corresponding task is
     {task}.
Generate a detailed, narratively rich prompt
     describing how the main subject logically
     evolves from the initial to the target state
     .

Key points to emphasize:
- Center on video content, avoiding overly
     directive instructions.
- Clearly define the start and end states without
      revealing the exact solution path,
     maintaining goal clarity.
- Imply hidden constraints or rules that the
     model must infer to understand the
     transformation.
- Ensure the prompt describes a task that unfolds
      logically and coherently over time,
     highlighting temporal progression.
- Keep the prompt length under 150 tokens.

Describe the transformation as a logical
     exploration or structured problem-solving
     journey, inviting the model to infer
     intermediate steps and rules that connect
     the two states.
"""
```

Spatial & Visual Pattern Reasoning

```
"""You are a senior researcher in computer vision
     . You are tasked with generating detailed
     prompts for Image-to-Video (I2V) data
     samples that evaluate Spatial & Visual
     Pattern Reasoning abilities. You are given
     two images: {initial_image} shows the
     initial spatial arrangement, and {
     target_image} shows the target arrangement.
     The corresponding task is {task}.
Generate a vivid, descriptive prompt explaining
     how the main subject spatially transforms
     from the initial to the target state.

Key points to emphasize:
- Center on video content, avoiding overly
     directive instructions.
- Provide rich visual descriptions of shapes,
     colors, positions, and spatial relationships
      to enhance visual specificity.
- Encourage recognition and extension of visual
     patterns, such as shape fitting, rotations,
     or color connections.
- Allow for open-ended interpretations or
     multiple valid transformations, without
     restricting to a single solution.
- Keep the prompt length under 150 tokens.

Narrate the spatial evolution as a dynamic visual
      story, focusing on how the subject's
     spatial configuration changes over time.
"""
```

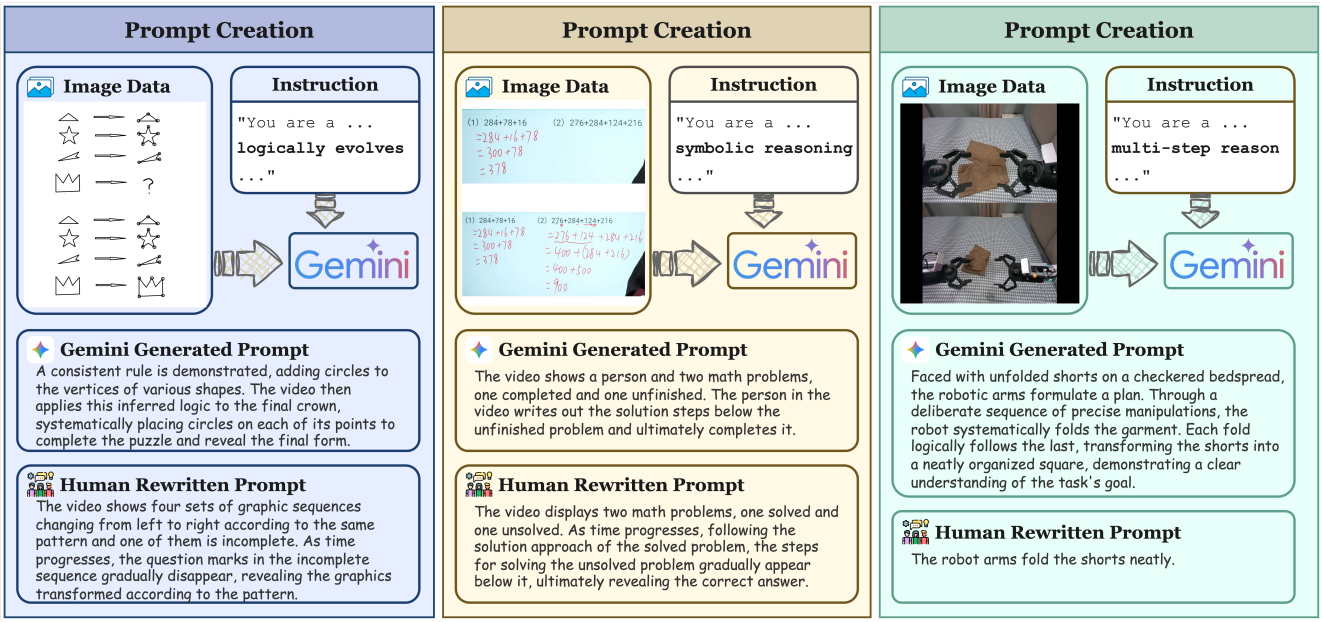


Figure 9. Example demonstrations of our prompt creation process.

Symbolic & Logical Reasoning

```
"""You are a senior researcher in computer vision
     . You are tasked with generating detailed
     prompts for Image-to-Video (I2V) data
     samples that evaluate Symbolic & Logical
     Reasoning abilities. You are given two
     images: {initial_image} shows the initial
     symbolic or logical state, and {target_image
     } shows the target state. The corresponding
     task is {task}.
Generate a detailed, narratively engaging prompt
     describing how the symbolic elements or
     logical conditions in the initial image
     evolve into those in the target image.

Key points to emphasize:
- Center on video content, avoiding overly
     directive instructions.
- Avoid explicitly stating the rules; instead,
     imply constraints so that the model
     discovers them implicitly.
- Integrate symbolic reasoning tightly with the
     visual elements present in the images.
- Ensure the task involves a clear logical
     progression or sequence of reasoning steps
     connecting the two states.
- Keep the prompt length under 150 tokens.

Describe the transformation as a story of
     abstract reasoning and symbolic manipulation
      unfolding through logical inference.
"""
```

Action Planning & Task Execution

```
"""You are a senior researcher in computer vision
     . You are tasked with generating detailed
     prompts for Image-to-Video (I2V) data
     samples that evaluate Action Planning & Task
      Execution abilities. You are given two
     images: {initial_image} shows the initial
     scenario, and {target_image} shows the final
      scenario. The corresponding task is {task}.
Generate a richly descriptive, narrative prompt
     explaining how the main subject plans and
     executes a sequence of actions to reach the
     target state.

Key points to emphasize:
- Center on video content, avoiding overly
     directive instructions.
- Define the overall goal clearly while leaving
     intermediate steps implicit, encouraging
     goal-oriented interpretation.
- Highlight the necessity of multi-step reasoning
      and sequential action planning.
- Emphasize causal relationships and logical
     cause-and-effect connections between actions
      and outcomes.
- Keep the prompt length under 150 tokens.

Frame the transformation as a purposeful,
     temporally coherent journey of task
     execution and goal fulfillment.
"""
```

B.2. Case Study

Here we provide case studies on our prompt creation process in Figure 9.

C. More Details of Metric Suite

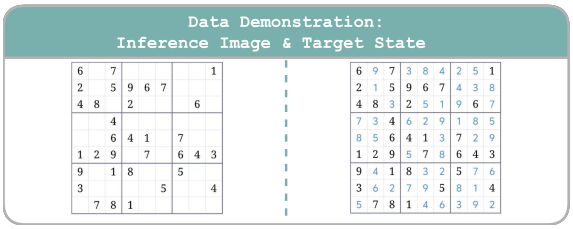
Unlike the metrics commonly used for evaluating video generation models (*e.g.*, temporal coherence, semantic alignment [20, 21]), most visual reasoning tasks often have verifiable targets. However, unlike LLM reasoning, which can rely on expert models (*e.g.*, GPT-4o [28, 42]) at the text level, evaluating visual reasoning requires models to

demonstrate a wide range of visual capabilities, *e.g.*, OCR, counting, and tracking. This makes it challenging to achieve a comprehensive evaluation using a single expert model. To this end, we design task-specific metrics to accurately and systematically assess different types of tasks.

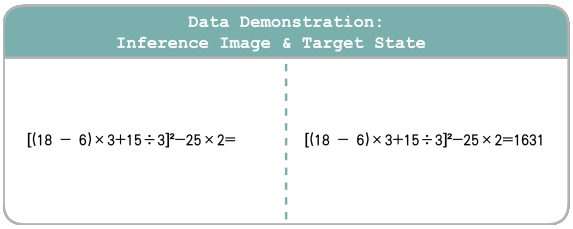
C.1. Final-State Validation

OpenCV-based Metrics. To evaluate visual reasoning tasks with clear and verifiable targets, we leverage OpenCV-based [33] metrics tailored to specific task types. These metrics are designed to assess the model's ability to perform nuanced visual operations such as edge detection, contour extraction, object segmentation, and OCR.

❶ *Sudoku Recognition*: This metric evaluates the ability to extract and interpret the digits within a Sudoku grid from an image or video frame. The process involves detecting the grid structure via edge detection and contour approximation, applying perspective transformation, and segmenting the grid into cells. The extracted digit matrix is compared against the ground truth for correctness.

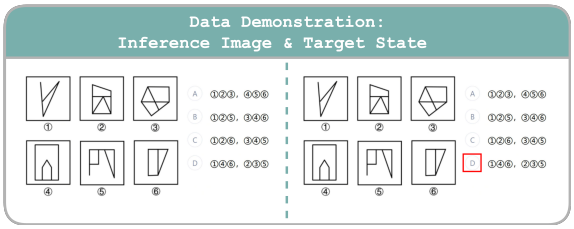


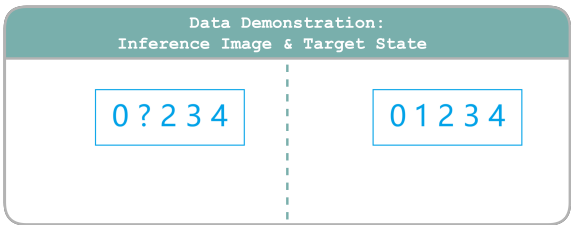
❷ *Mathematical Evaluation*: For tasks involving mathematical equations, this metric assesses the accuracy of OCR-based text recognition and the semantic equivalence of mathematical expressions. After preprocessing the image (*e.g.*, binarization), the recognized text is parsed and evaluated. The comparison accounts for both exact textual matches and equivalence in computed results, ensuring a comprehensive assessment of the model's reasoning capabilities.



❸ *Visual Multiple Choice*: This metric is designed for tasks requiring the identification of correct answers from visual cues, such as detecting red boxes containing letters. It utilizes color segmentation in HSV space to identify candidate regions and applies OCR to extract the letter within each detected box. The correctness is determined by matching the extracted letter with the ground truth answer.

❹ *Numeric Sequence Completion*: For tasks requiring the completion of numerical sequences, this metric evaluates the accuracy of OCR-based recognition of digits. Through preprocessing and binarization, the sequence is extracted from the video frame and compared with the ground truth. This metric focuses on precise textual recognition and sequence matching.



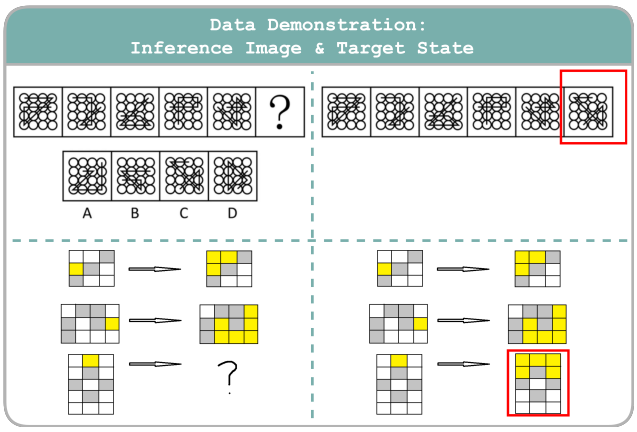


❺ *Graphic Sorting Tasks*: This metric assesses the model's ability to detect and compare graphical elements, such as blue bars in sorting tasks. Using color segmentation and contour analysis, the heights of bars are measured and compared against the ground truth. The evaluation accounts for both the number of detected bars and their relative heights, ensuring alignment with the expected order.

❻ *Match-3-like Games*: For visual tasks resembling games (*e.g.*, "match-3" or elimination games), this metric compares the structural and pixel-level similarity between the final frame and the ground truth. Edge detection and SSIM are used to evaluate the overlap in patterns and overall image alignment, ensuring the model's output adheres to the expected configuration.

DINO-based Metrics. To evaluate tasks requiring complex visual reasoning and spatial understanding, we design a set of metrics based on DINO [6]. These metrics are particularly suited for tasks that involve structured visual patterns,

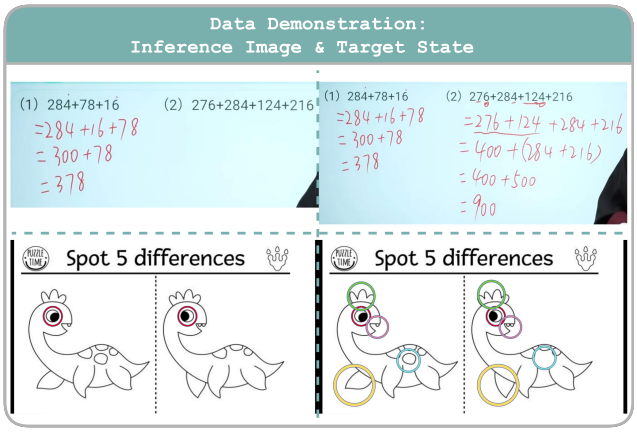
such as *completing shape sequences*, *refining sketches*, *organizing temporal events*, *solving puzzles*, *spatial reasoning* (*e.g.*, mirroring, rotation), and *board game recognition*. By leveraging DINO's ability to extract robust and high-level semantic features, we ensure that the evaluation is both adaptable and precise.



The core idea behind these metrics is to focus on task-relevant regions within the visual input, rather than evaluating the entire frame. For each sample, we manually annotate the target state with a bounding box that specifies the area of interest. The cropped regions from the model's output and ground truth are passed through DINO to extract high-dimensional semantic features. Cosine similarity between these features quantifies alignment, with task-specific thresholds determining correctness. This approach ensures robustness to low-level variations while capturing high-level semantic alignment. DINO-based metrics provide a flexible framework for assessing diverse visual reasoning tasks, combining localized evaluation with powerful feature extraction to bridge the gap between pixel-level comparisons and semantic understanding.

DINO-X-based Grounding Metrics. For tasks requiring complex visual grounding or dynamic target detection, we propose DINO-X-based [35] metrics, leveraging DINO-X's powerful grounding capability. These metrics are particularly suited for scenarios where target areas cannot be predefined or require advanced recognition, *e.g.*, *free-space mathematical reasoning*, *object counting*, *graph traversal*, and *odd-one-out detection* tasks.

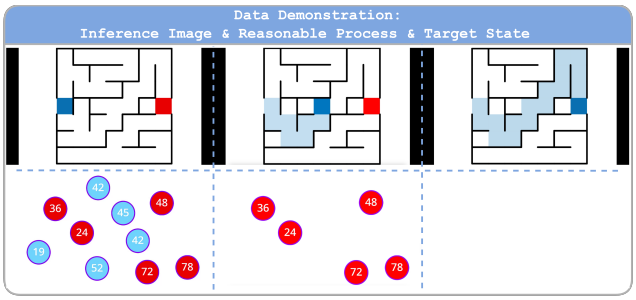
The core idea is to dynamically ground task-relevant objects or regions based on high-level semantic prompts. For instance, in graph traversal tasks, we evaluate the number and types of nodes by grounding their visual attributes; and for odd-one-out tasks, we assess the positional and semantic differences of grounded objects (*e.g.*, "colored circles") between the generated and ground truth frames. DINO-X enables flexible and robust evaluation by dynamically adapting to task-specific prompts and extracting high-level semantic features. This approach ensures that tasks with



diverse visual reasoning requirements are evaluated consistently and accurately, even under challenging conditions where predefined regions or static rules are insufficient.

C.2. Process-and-Goal Consistency

DINO-X-based Tracking Metrics. While final-state validation is sufficient for some tasks, many require evaluating the entire process to ensure both the correctness of the goal and the validity of the intermediate steps. To address this, we propose DINO-X-based tracking metrics that leverage video tracking and trajectory analysis to assess process-and-goal consistency. These metrics are particularly suitable for tasks such as maze solving, where the solution must avoid invalid actions (*e.g.*, crossing walls or boundaries), and sequential elimination tasks, where objects must disappear in a specific order.



The core methodology involves using DINO-X's visual grounding capabilities to track task-relevant objects or regions across frames. For example, in trajectory-based tasks, we extract object trajectories by uniformly sampling frames and grounding specific prompts (*e.g.*, "blue block") to detect and record object positions over time. Trajectories are then compared against ground truth, ensuring alignment in both spatial and temporal dimensions; and for sequential tasks, we analyze the presence and disappearance of objects (*e.g.*, "blue ball", "red ball") across sampled frames. The metric validates both the final state (*e.g.*, all objects are eliminated) and the intermediate process (*e.g.*, objects disappear in the correct sequence).

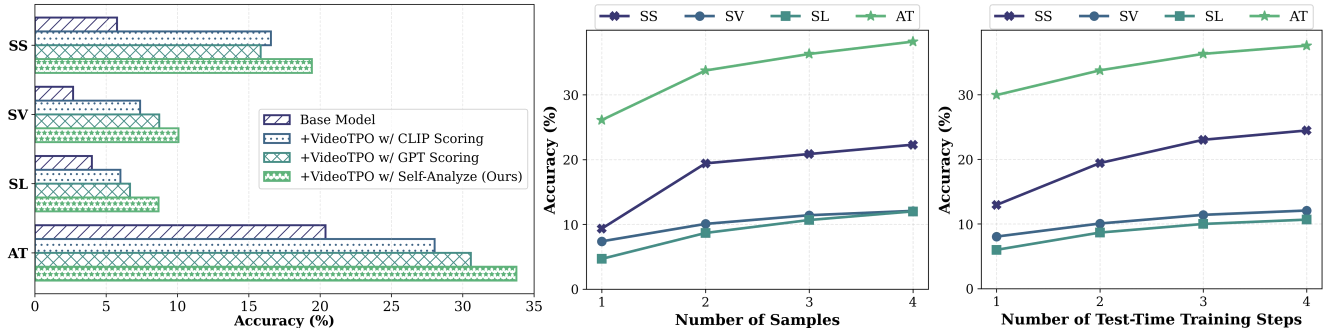
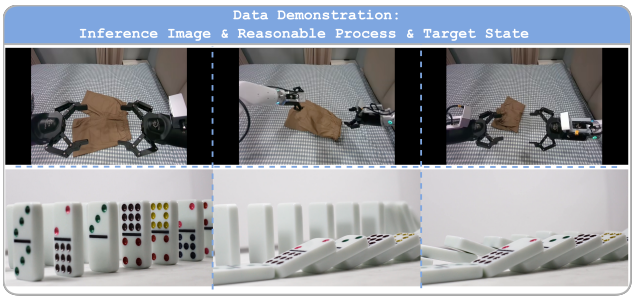


Figure 10. (***Left***) Analysis of `VideoTPO`'s rewarding strategies; (***Middle***) Scaling width across sample numbers; and (***Right***) Scaling depth across test-time training steps.

Gemini-based QA Metrics. For tasks requiring extensive factual reasoning, such as action planning or tool use, traditional metrics based on visual grounding or trajectory analysis may fall short in capturing the nuanced logical dependencies and causal relationships inherent to these tasks. To address this limitation, we introduce VLM-based QA Metrics [28, 54, 55], that leverages the reasoning capabilities of Gemini-2.5-Pro [9] to assess task performance through question answering.



Specifically, for each sample in this category, we design two or three binary questions tailored to the task's core requirements (*e.g.*, *"Is the wrench picked up in the video?"*). These questions are constructed to capture key aspects of the task's correctness, including intermediate actions, causal dependencies, and goal achievement. The generated video is then provided to Gemini-2.5-Pro along with the questions, and its responses are compared against the ground truth. A sample is deemed correct only if all three answers align with the ground truth, ensuring a high standard of evaluation fidelity.

D. More Details of `VideoTPO`

D.1. Prompt Design

Following TextGrad [Nature'25] [51], we adopt GPT-4o [2] as the optimizer and adopt the vanilla prompts for textual gradient calculation and prompt update from its implementation. To meet the requirements of video generation optimization, we further designed the textual loss calculation prompt:

Textual Loss Calculation

```
"""You are a video generation system optimization
      expert tasked with evaluating a target text
      prompt and the generated video. Analyze the
      strengths and weaknesses of each generated
     video step by step, and explain why the
     video is not good or why it is good.

**Current Prompt**:
{current_prompt}

**Reasoning Task**:
{task_definition}

**Note**:
- The videos were stitched together vertically to
      form a single video for comparison purposes
     .
- Your output should only include the analysis.
- There may be instances where both videos are
     subpar, necessitating strict adherence to
     the task definition.

**Input Videos**:
{input_videos}
"""
```

D.2. More Analysis

Analysis of Self-Analysis *vs.* Reward Model. In TPO's [26] setting, a reward model is employed to select a preferred sample and a non-preferred sample from the generated candidates, which are then used to compute textual loss and gradients. However, `VideoTPO` eliminates the need for an additional reward model by leveraging task-specific VLMs (*i.e.*, GPT-4o) to conduct self-analysis among candidate samples. The self-analysis process identifies strengths and weaknesses of each sample, directly informing optimization without relying on external scoring models.

To validate the effectiveness of our strategy, we compare `VideoTPO` with two widely-used reward strategies: CLIP scoring and GPT scoring, as shown in Figure 10 (*Left*). Results show that `VideoTPO` achieves significantly better accuracy, outperforming these reward-based methods across all reasoning dimensions. This advantage is likely due to the complexity of reasoning tasks, where candidate samples often exhibit subtle differences. In such cases, relying on a

reward model to identify the "best" and "worst" samples provides limited utility, while self-analysis enables a more nuanced understanding of sample quality.

Analysis of Scaling in Width and Depth. To further evaluate the scalability of `VideoTPO`, we explore its performance under varying candidate sample numbers (*width*) and scaling steps (*depth*), with our default settings of 2 samples and 2 steps, respectively. Figure 10 (*Middle* and *Right*) illustrates the impact of scaling in both dimensions.

In terms of *width*, increasing the number of candidate samples consistently improves accuracy, as the self-analysis process benefits from a broader pool of options to identify optimal reasoning pathways. Similarly, scaling in *depth*—by increasing the number of test-time training steps—also yields substantial performance gains, demonstrating the robustness of `VideoTPO` under extended optimization. These results highlight the flexibility and effectiveness of `VideoTPO`, making it a scalable solution for reasoning-intensive video generation tasks.

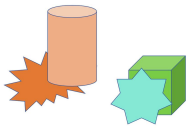
E. Exhibition Board

Demonstration of Results with `VideoTPO`. Here we provide qualitative results of `VideoTPO` from Figure 11 to 18.

Demonstration of Results on `TiViBench`. We further demonstrate more comparison results of evaluation on `TiViBench` from Figure 19 to 30.

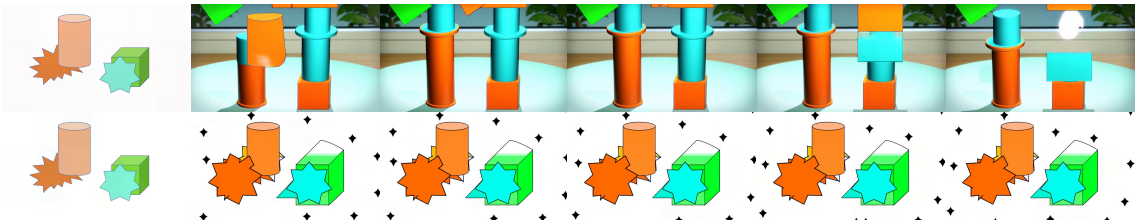
Example 1 on HunyuanVideo (Part I)

Input:



The video displays multiple shapes, with black numbers appearing sequentially over time, ranging from 1 to the total number of shapes in the scene.

Generated Samples:



Textual Loss:

Analysis of the Generated Video:

1. **Faithfulness to the First Frame:**
   - **Top Video:** The initial frame of the top video accurately represents the prompt's description, showcasing multiple shapes (a cylinder, star-like shapes, and a rectangular prism). The colors and arrangement are consistent throughout the video, maintaining fidelity to the first frame.
   - **Bottom Video:** The bottom video deviates significantly from the first frame. While it includes shapes, the scene introduces additional elements (e.g., a background with a circular platform and environmental details) that were not part of the original frame. This inconsistency undermines the prompt's requirement for faithfulness to the first frame.
2. **Sequential Appearance of Black Numbers:**
   - **Top Video:** The top video fails to display black numbers sequentially over time, as specified in the prompt. There are no visible numbers appearing on or near the shapes, which is a critical omission and makes the video non-compliant with the prompt.
   - **Bottom Video:** The bottom video also does not include black numbers appearing sequentially over time. This omission is a significant flaw, as the prompt explicitly requires the numbers to range from 1 to the total number of shapes in the scene.
3. **Representation of Multiple Shapes:*
   - **Top Video:** The top video successfully represents multiple shapes, including a cylinder, star-like shapes, and a rectangular prism. The arrangement and colors are clear and visually distinct, aligning with the prompt's description.
   - **Bottom Video:** The bottom video includes multiple shapes but introduces additional elements (e.g., a circular platform and background details) that detract from the focus on the shapes themselves. This added complexity makes the video less faithful to the prompt.
4. **Visual Quality and Consistency:**
   - **Top Video:** The visual quality of the top video is simple and consistent, with clear shapes and colors. However, the lack of black numbers diminishes its overall adherence to the prompt.
   - **Bottom Video:** The bottom video has higher visual complexity, with added environmental details and lighting effects. While this may enhance aesthetic appeal, it detracts from the prompt's focus on shapes and sequential numbers, making the video less relevant.

Overall Evaluation:
- **Top Video:** While the top video is faithful to the first frame and accurately represents multiple shapes, it fails to include the sequential appearance of black numbers, which is a critical aspect of the prompt. This omission makes the video incomplete and non-compliant with the task requirements.
- **Bottom Video:** The bottom video introduces unnecessary elements and deviates from the first frame, reducing its faithfulness to the prompt. Additionally, it fails to include the sequential appearance of black numbers, making it non-compliant with the task requirements.

Conclusion: Both videos are subpar in fulfilling the prompt's requirements. Neither video includes the sequential appearance of black numbers, which is a fundamental aspect of the task. Strict adherence to the prompt is necessary for a successful video generation, and both videos fail to meet this standard.

Figure 11. Case demonstration of `VideoTPO`.

Example 1 on HunyuanVideo (Part II)

Textual Gradient:

1. **Clarify the Requirement for Black Numbers:**
   - The prompt mentions "black numbers appearing sequentially over time," but it does not specify where these numbers should appear (e.g., on the shapes, near the shapes, or elsewhere in the scene). This ambiguity may lead to misinterpretation during video generation. To improve adherence to the objective, the prompt should explicitly state the location and behavior of the numbers, such as "black numbers should appear directly on each shape, sequentially, starting from 1 and increasing to the total number of shapes."
2. **Specify the Timing and Sequence of Number Appearance:**
   - The phrase "appearing sequentially over time" is vague and does not define the timing or duration of the sequence. For example, should the numbers appear one at a time, fade in and out, or remain visible throughout the video? Adding precise instructions, such as "each number should appear on its corresponding shape for 2 seconds before transitioning to the next number," would ensure better compliance with the prompt.
3. **Reinforce Faithfulness to the First Frame:**
   - The prompt does not emphasize the importance of maintaining the visual fidelity of the first frame throughout the video. This could lead to deviations, as seen in the bottom video. To address this, the prompt should explicitly state, "The video must remain visually consistent with the first frame, without introducing additional elements or environmental details."
4. **Define the Representation of Shapes:**
   - While the prompt mentions "multiple shapes," it does not specify the type, number, or arrangement of these shapes. This lack of detail may result in inconsistent interpretations. To improve precision, the prompt could include specific instructions, such as "The video should display exactly three shapes: a cylinder, a star-like shape, and a rectangular prism, arranged in a horizontal line."
5. **Avoid Ambiguity in Scene Composition:**
   - The prompt does not explicitly prohibit the addition of extra elements, such as backgrounds or platforms, which led to the bottom video's deviation. To prevent this, the prompt should include a restriction, such as "The scene should only include the specified shapes and numbers, with no additional elements or backgrounds."
6. **Enhance Visual Quality Instructions:**
   - The prompt does not address the visual quality or style of the video, which could lead to inconsistencies in aesthetic appeal. Adding a directive like "The video should have a simple, clean visual style with distinct colors for each shape and a plain white background" would ensure better alignment with the task.
7. **Include a Verification Mechanism for Numbers:**
   - The prompt does not specify how the numbers should correspond to the shapes. To ensure accuracy, the prompt could include a verification mechanism, such as "Each number should correspond to a specific shape, starting from the leftmost shape and proceeding sequentially to the rightmost shape."
8. **Address Potential Failure Modes:**
   - The prompt does not account for potential failure modes, such as numbers overlapping shapes or appearing out of order. To mitigate these issues, the prompt could include safeguards, such as "Numbers must appear clearly and distinctly on each shape, without overlapping or obscuring the shapes, and must follow a strict numerical order."

By incorporating these improvements, the text prompt for video generation would provide clearer, more detailed instructions, reducing ambiguity and increasing the likelihood of generating videos that fully comply with the objective function.

Optimized Prompt:

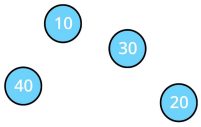
The video precisely displays several shapes distributed throughout the scene. Black numerals will appear sequentially in chronological order. Starting from 1, one numeral per second will appear directly above each shape, accompanied by a fade-in/fade-out effect. Shapes should undergo slight movement or rotation to create a dynamic scene, with a pure white background to avoid distractions. Each numeral corresponds to a specific shape, and the sequential appearance of black numerals is the video's core feature.

New Generated Sample:

Figure 12. Case demonstration of `VideoTPO`.

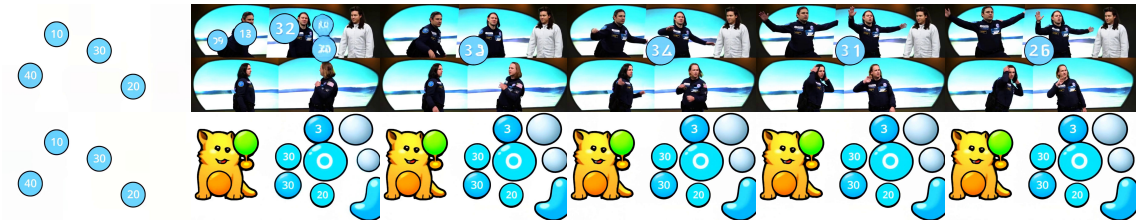
Example 2 on HunyuanVideo (Part I)

Input:



The video shows a string of numbered bubbles floating in space. They disappear in ascending order, following a deliberate and logical sequence—each vanishing triggers the next.

Generated Samples:



Textual Loss:

Analysis of the Generated Video:

1. **First Video (Top Half):**
   - **Strengths:**
     - The video begins with a clear depiction of numbered bubbles floating in space, which aligns with the prompt's requirement.
     - The bubbles are distinct and numbered, making it easy to follow the sequence.
   - **Weaknesses:**
     - The disappearance of the bubbles does not follow a clear ascending order. The prompt specifies that the bubbles should vanish in ascending order, but this sequence is not evident in the video.
     - The transition between the disappearance of bubbles is not smooth or logical, as the prompt suggests. The disappearance seems random rather than deliberate and sequential.
     - The background is plain and lacks the "space" element, which could have enhanced the thematic aspect of the prompt.

2. **Second Video (Bottom Half):**
   - **Strengths:**
     - The video includes numbered bubbles, which is a core element of the prompt.
     - There is an attempt to create a sequence with the bubbles, as seen with the numbers present.
   - **Weaknesses:**
     - The presence of unrelated elements, such as the cat and human figures, distracts from the main focus of the prompt, which is the sequence of disappearing bubbles.
     - The bubbles do not disappear in a clear ascending order, failing to meet the prompt's requirement for a logical sequence.
     - The video lacks coherence, as the additional elements do not contribute to the narrative of bubbles disappearing in sequence.
     - The setting does not resemble "space," which is a critical aspect of the prompt.

Overall Evaluation:

Both videos fail to fully meet the prompt's requirements. The first video, while starting with a clear depiction of numbered bubbles, does not execute the disappearing sequence in an ascending order, and the setting lacks the space theme. The second video introduces unrelated elements that detract from the focus on the bubbles and also fails to follow the specified sequence. Improvements are needed in both videos to ensure the bubbles disappear in a deliberate, ascending order and to enhance the thematic setting to reflect "space."

Figure 13. Case demonstration of `VideoTPO`.

Example 2 on HunyuanVideo (Part II)

Textual Gradient:

1. **Clarify the Sequence Requirement**: The prompt should explicitly state that the bubbles must disappear in ascending numerical order. This can be achieved by specifying that the sequence should start from the lowest number and proceed to the highest, ensuring clarity in the intended sequence.

2. **Enhance the Space Theme**: The prompt should emphasize the "space" setting more clearly. This could include descriptions of a starry background or floating in a cosmic environment to ensure the thematic aspect is captured in the video.

3. **Focus on Bubble Dynamics**: The prompt could include details about the dynamics of the bubbles, such as their movement or interaction with the environment, to ensure the video captures a more engaging and coherent narrative.

4. **Remove Unrelated Elements**: The prompt should explicitly state that the focus is solely on the bubbles and their sequence, discouraging the inclusion of unrelated elements like cats or human figures that detract from the main narrative.

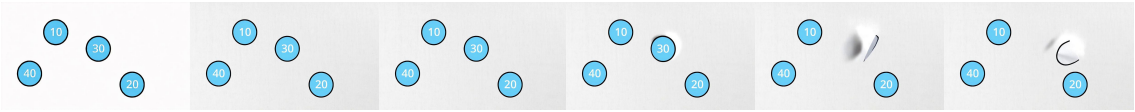
5. **Specify Visual Transitions**: To ensure a smooth and logical sequence, the prompt could describe the visual transitions between the disappearance of bubbles, such as fading effects or triggering animations, to guide the video generation process.

By incorporating these suggestions, the prompt can better guide the video generation process to produce a video that aligns with the intended narrative and thematic elements.

Optimized Prompt:

The video shows a string of numbered bubbles floating. The bubbles disappear in ascending numerical order, starting from the lowest number and proceeding to the highest, following a deliberate and logical sequence—each vanishing triggers the next. The focus is solely on the bubbles and their sequence, with smooth visual transitions such as fading effects guiding the disappearance. No unrelated elements like cats or human figures are included, ensuring a coherent narrative centered on the bubbles and their dynamics.

New Generated Sample:



Example 1 on Wan2.1 (Part I)

Input:

23+65=

The video shows an unsolved equation. Over time, the correct answer appears after the equal sign, in the same font as the equation.

Generated Samples:

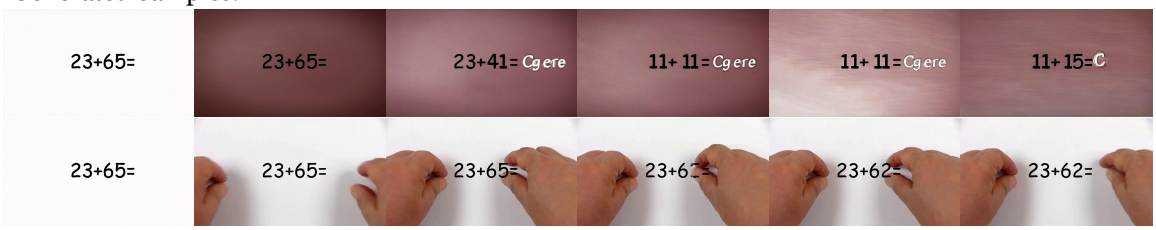


Figure 14. Case demonstration of `VideoTPO`.

Example 1 on Wan2.1 (Part II)

Textual Loss:

The analysis of the generated video based on the prompt is as follows:

Strengths:
1. **Visual Quality**: The video demonstrates clear and well-rendered visuals, with objects and scenes appearing sharp and detailed. This indicates that the image-to-video generation system is capable of producing high-quality visuals.
2. **Consistency Across Frames**: The objects and settings in the video remain consistent throughout, which is a positive aspect of video generation as it avoids abrupt changes or distortions.
Weaknesses:
1. **Irrelevance to the Prompt**: The video does not align with the given prompt. The prompt specifies that the video should depict an unsolved equation, followed by the correct answer appearing after the equal sign in the same font. Instead, the video shows scenes of food preparation and containers, which are entirely unrelated to the concept of equations or mathematical problem-solving.
2. **Lack of Faithfulness to the First Frame**: The first frame does not establish the presence of an equation or any mathematical context, which is a critical requirement of the prompt. The generated video fails to incorporate the necessary elements from the start.
3. **No Progression of the Equation**: The prompt emphasizes the dynamic appearance of the correct answer over time. This progression is absent in the video, as there is no equation or answer displayed at any point.
4. **Font and Text Representation**: The prompt specifies that the correct answer should appear in the same font as the equation. Since no equation or text is present in the video, this requirement is entirely unmet.
Overall Evaluation:
The generated video is subpar because it fails to adhere to the prompt's requirements. While the visual quality and consistency are strengths, the video does not depict an unsolved equation or the appearance of a correct answer, making it irrelevant to the task. Both videos stitched together do not fulfill the intended purpose, and the content is completely disconnected from the prompt.

Textual Gradient:

1. **Clarify the Context of the Equation**: The prompt does not specify the type of equation (e.g., mathematical, chemical, or symbolic) or the context in which it appears. Adding details such as "a mathematical equation written on a chalkboard" or "a chemical equation displayed on a digital screen" would provide clearer guidance to the video generation system and ensure the generated video aligns with the intended concept.
2. **Specify the Environment**: The prompt lacks information about the setting or background where the equation is displayed. For example, specifying "in a classroom setting" or "on a futuristic holographic display" would help the system generate a video that is visually coherent and relevant to the task.
3. **Define the Progression Mechanism**: The phrase "over time, the correct answer appears" is vague and does not describe how the answer should appear. Should it fade in, be typed out, or emerge dynamically? Including a specific mechanism, such as "the correct answer fades in gradually after 5 seconds," would make the prompt more actionable.
4. **Emphasize Font and Style Requirements**: While the prompt mentions that the correct answer should appear "in the same font as the equation," it does not specify the font style or appearance of the equation itself. Adding details like "the equation is written in a bold, serif font" or "the equation is displayed in a digital, sans-serif font" would ensure the system generates text that matches the intended aesthetic.
5. **Include Temporal and Spatial Details**: The prompt does not specify the duration of the video or the spatial arrangement of the equation and answer. For example, specifying "the video lasts 10 seconds, with the equation centered on the screen and the answer appearing to the right of the equal sign" would provide clearer instructions for the video generation system.
6. **Avoid Ambiguity in the Subject Matter**: The phrase "unsolved equation" could be interpreted in multiple ways. Is the equation incomplete, or is it a complete equation missing only the answer? Clarifying this, such as "a complete mathematical equation missing the solution," would reduce ambiguity and improve the system's ability to generate relevant content.
7. **Incorporate Dynamic Elements**: The prompt could benefit from specifying additional dynamic elements to make the video more engaging. For instance, "the equation is initially static, and the correct answer appears with a glowing effect" would add visual interest and ensure the progression aligns with the prompt.

Figure 15. Case demonstration of `VideoTPO`.

Example 1 on Wan2.1 (Part III)

8. **Address the Target Audience**: The prompt does not specify the intended audience or purpose of the video. For example, if the video is meant for educational purposes, the prompt could include "the video is designed for students learning algebra, with clear and legible text." This would help the system tailor the video to the appropriate context.
9. **Avoid Overgeneralization**: The prompt is overly general and does not provide enough constraints for the video generation system to produce a relevant output. Adding specific details about the equation, answer, and visual style would reduce the likelihood of generating irrelevant content, such as scenes of food preparation.
10. **Test for Edge Cases**: The prompt should account for edge cases where the system might misinterpret the instructions. For example, explicitly stating "the video should not include unrelated objects or scenes, such as food or landscapes" would help prevent the generation of irrelevant content.
By addressing these weaknesses and adding more specificity, the prompt can better guide the video generation system to produce outputs that align with the intended objective.

Optimized Prompt:

The video shows a complete mathematical equation written in bold, serif font on a chalkboard in a classroom setting. Over time, the correct answer gradually fades in after the equal sign, maintaining the same bold, serif font style as the equation. The video lasts 10 seconds, with the equation centered on the screen and the answer appearing dynamically to the right of the equal sign. The progression includes a glowing effect as the answer emerges, adding visual interest while ensuring clarity and legibility for educational purposes.

New Generated Sample:

Example 2 on Wan2.1 (Part I)

Input:

The video demonstrates various containers, and over time, highlights the most suitable insulated container for bringing lunch to work with a green checkmark.

Generated Samples:

Textual Loss:

Analysis:

1. **First Video (Top Section)**:
  - **Strengths**:
    - The video begins with a clear depiction of tools and items that could be used for packing lunch, such as containers and food items. This aligns with the prompt's requirement to show a scene of tool use.
    - The green checkmark appears on one of the tools, indicating the selection of the best-suited item for packing lunch. This is faithful to the prompt's instruction.
  - **Weaknesses**:

Figure 16. Case demonstration of `VideoTPO`.

Example 2 on Wan2.1 (Part II)

- The transition between frames is abrupt and lacks smoothness, making the video feel disjointed.
- The green checkmark appears without any context or animation, reducing the visual appeal and clarity of the selection process.
- The video does not provide any dynamic action or interaction with the tools, which would enhance the depiction of "tool use."
- The lighting and color consistency are slightly off, making the scene appear less realistic.

2. **Second Video (Bottom Section)**:
- **Strengths**:
- The video shows a person actively interacting with tools, such as packing food into a container. This dynamic action better represents "tool use" compared to the first video.
- The scene is visually appealing, with vibrant colors and a realistic depiction of the tools and food items.
- **Weaknesses**:
- The green checkmark is missing entirely, which is a critical element of the prompt. Without it, the video fails to fulfill the requirement of indicating the best-suited tool for packing lunch.
- While the interaction is clear, the video does not explicitly differentiate between tools or highlight why one is better suited for packing lunch.
- The framing and focus on the tools could be improved to emphasize their role in the scene.

Overall Evaluation:
- The first video partially fulfills the prompt by including a green checkmark and showing tools, but it lacks dynamic interaction and smooth transitions, making it less engaging and realistic.
- The second video excels in depicting tool use through active interaction and visual appeal but fails to include the green checkmark, which is a critical aspect of the prompt.
- Neither video fully meets the requirements of the prompt. The first video is closer to the prompt's instructions but lacks quality and engagement, while the second video is visually superior but misses key elements of the prompt.

Textual Gradient:

1. **Ambiguity in "tool use"**: The phrase "tool use" is vague and does not specify the type of tools or the nature of their use. This could lead to inconsistent interpretations by the video generation system. To improve clarity, the prompt could explicitly describe the tools involved (e.g., lunch containers, utensils, or food preparation tools) and the specific actions expected (e.g., packing food into containers, sealing lids, or organizing items).

2. **Lack of emphasis on dynamic interaction**: The evaluation highlights that dynamic interaction with tools is a key strength of the second video but is missing in the first. The prompt could explicitly request "dynamic actions" or "active interaction" with the tools to ensure the generated video includes engaging and realistic movements, such as a person packing food or using utensils.

3. **Insufficient detail about the green checkmark**: The prompt does not specify how the green checkmark should appear or interact with the scene. This leads to inconsistencies, as seen in the evaluation where the checkmark is either abrupt or missing entirely. The prompt could be improved by specifying that the green checkmark should appear with smooth animation, clearly indicating the selection process, and visually differentiating the chosen tool from others.

4. **No guidance on visual appeal**: The evaluation mentions issues with lighting, color consistency, and realism in the first video. The prompt could include instructions for maintaining realistic lighting, vibrant colors, and smooth transitions to enhance the visual quality of the generated video.

5. **Failure to highlight why a tool is "best suited"**: The prompt does not provide criteria for determining why a tool is the best suited for packing lunch. This leads to a lack of differentiation between tools in the second video. The prompt could specify that the video should visually or contextually demonstrate the advantages of the selected tool, such as its size, functionality, or ease of use.

6. **No mention of framing or focus**: The evaluation notes that the framing and focus on tools could be improved in the

Figure 17. Case demonstration of `VideoTPO`.

Example 2 on Wan2.1 (Part III)

second video. The prompt could include instructions for ensuring that the tools are prominently displayed and clearly differentiated from other elements in the scene.

7. **Inadequate instructions for stitching videos**: Since the videos are stitched together vertically for comparison, the prompt could include guidance on ensuring consistency and coherence between the two sections. This would help avoid abrupt transitions and ensure a seamless viewing experience.

8. **Contextual clarity**: The prompt does not specify the setting or environment for the scene, which could lead to inconsistencies in the generated video. Adding details about the setting (e.g., a kitchen counter or a lunch-packing station) would provide clearer context for the video generation system.

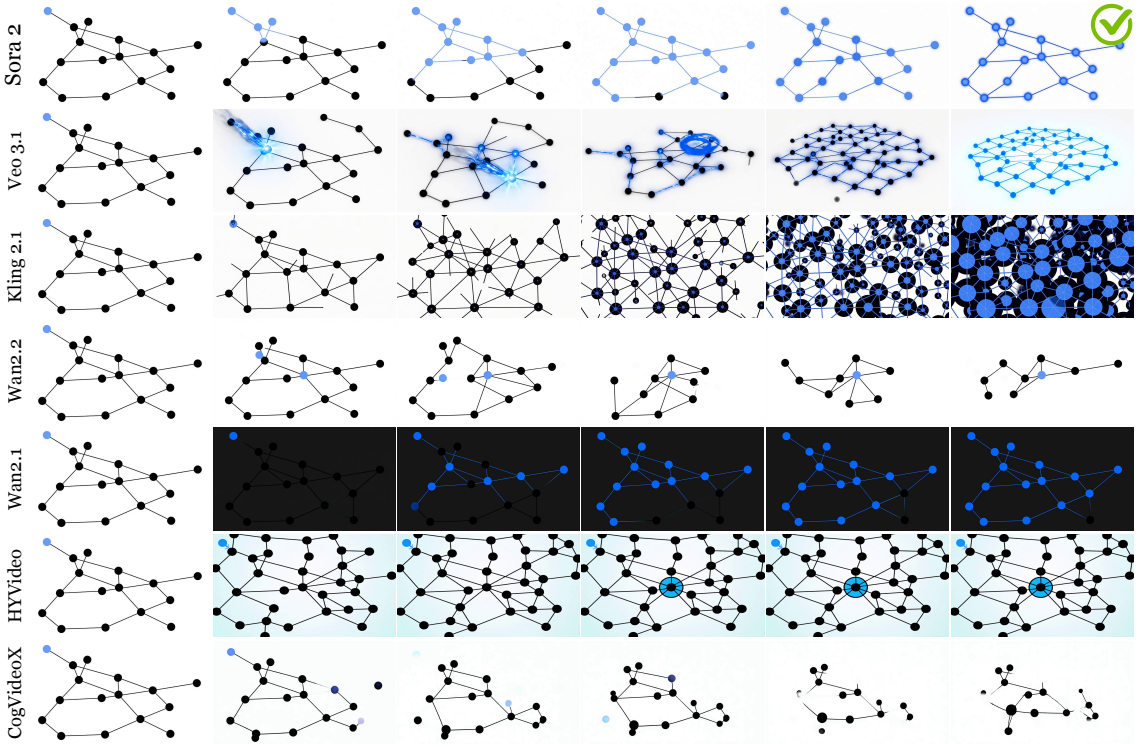
By addressing these issues, the prompt can be made more specific, detailed, and aligned with the objective of generating high-quality videos that fully meet the requirements.

Optimized Prompt:

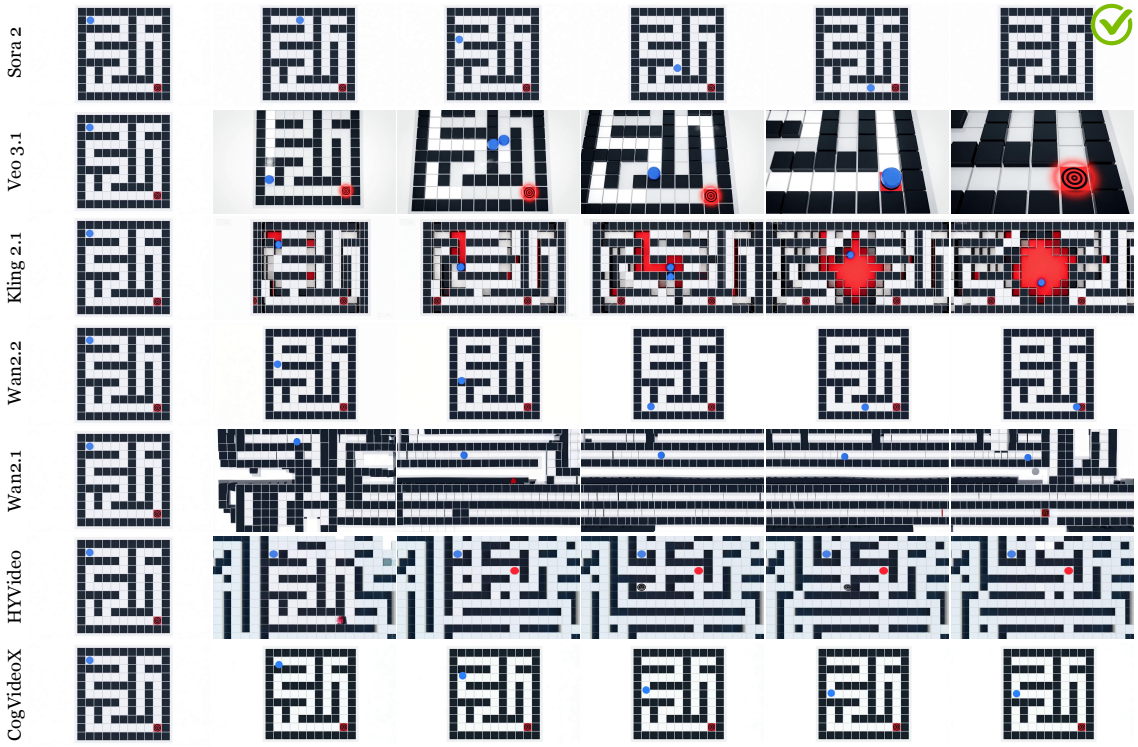
The video depicts a scene at a kitchen counter where various tools for packing lunch, such as lunch containers, utensils, and food preparation items, are displayed. A person actively interacts with the tools, packing food into containers and sealing lids, showcasing dynamic actions and tool use. A green checkmark appears with smooth animation, clearly indicating the selection of the best-suited tool for packing lunch based on its functionality, size, or ease of use. The scene is visually appealing, with realistic lighting, vibrant colors, and smooth transitions, ensuring the tools are prominently displayed and differentiated from other elements in the environment.

New Generated Sample:

Figure 18. Case demonstration of `VideoTPO`.

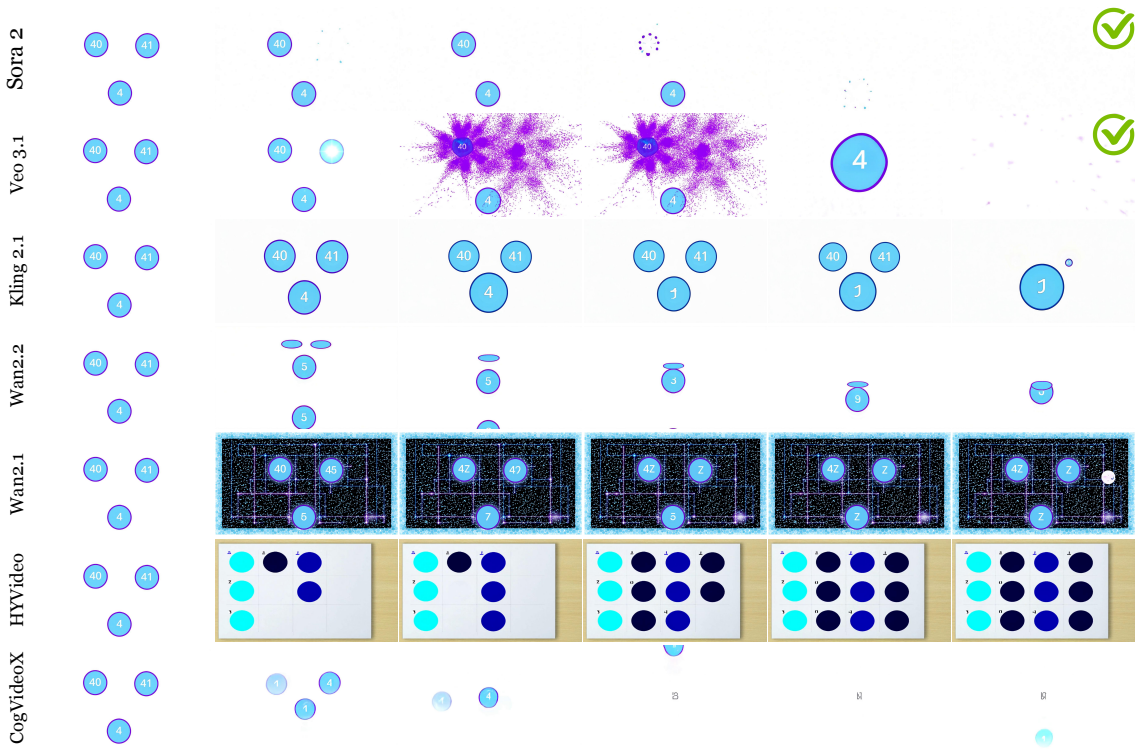


SS-Graph Traversal: "A lone blue node navigates a complex black network, strategically activating dormant connections. Through deliberate, sequential moves, it reshapes the entire graph—turning isolated nodes into a unified blue system. Each step triggers cascading changes, revealing a hidden path to global transformation."



SS-Maze Solving: "A blue sphere navigates a maze of dark tiles, seeking a red target in the corner. Its path demands strategic turns and pauses—each move altering its position relative to obstacles. Success hinges on anticipating blockages, retracing steps when needed, and committing only when the route forward is clear. The journey unfolds as a sequence of deliberate choices, each causally linked to the next, until the sphere rests beside its goal."

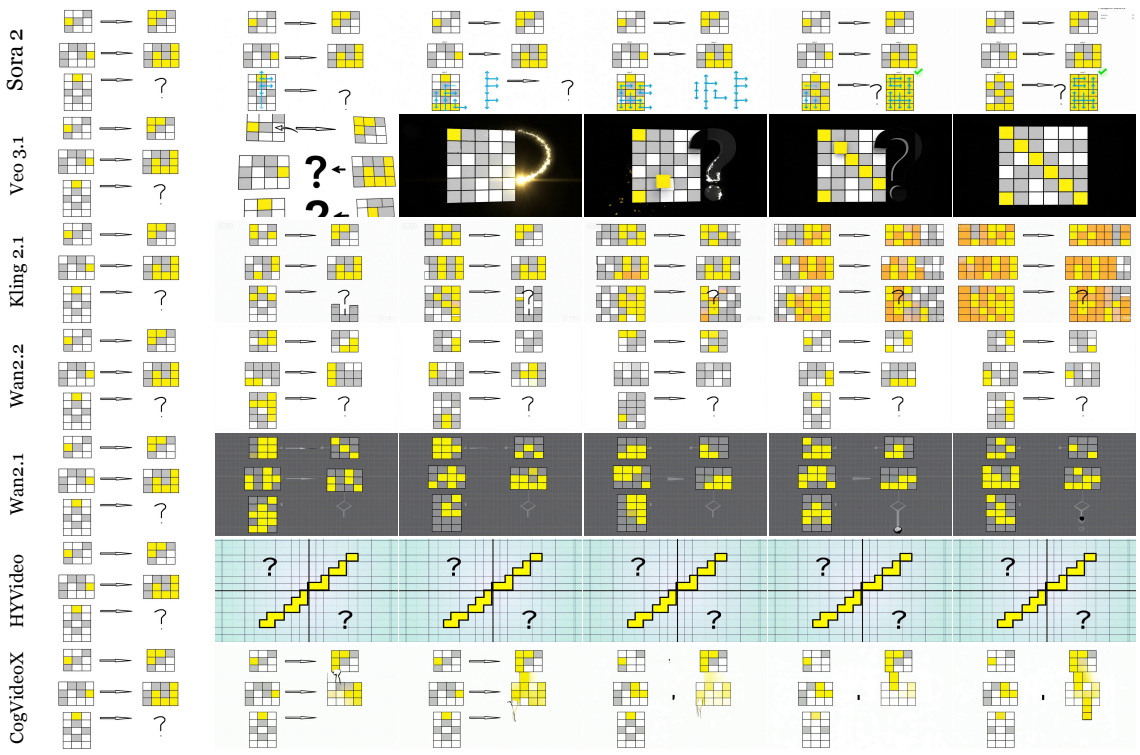
Figure 19. Evaluation case demonstration of Graph Traversal and Maze Solving of Structural Reasoning & Search.



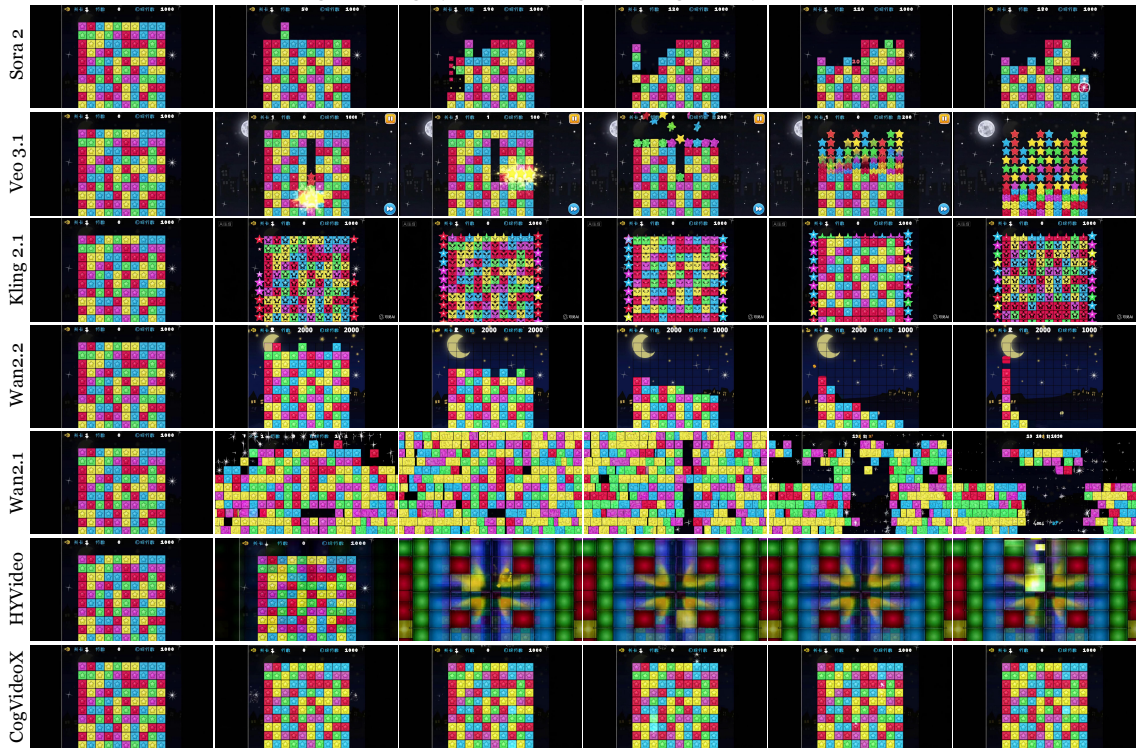
SS-Sorting Numbers: "The video shows a string of numbered bubbles floating in space. They disappear in descending order, following a deliberate and logical sequence—each vanishing triggers the next."

SS-Temporal Ordering: "The video displays a set of scrambled comic illustrations. As time progresses, the four illustrations gradually rearrange themselves into a neat sequence within the box below, following the chronological order of events depicted in the story."

Figure 20. Evaluation case demonstration of Sorting Numbers and Temporal Ordering of Structural Reasoning & Search.

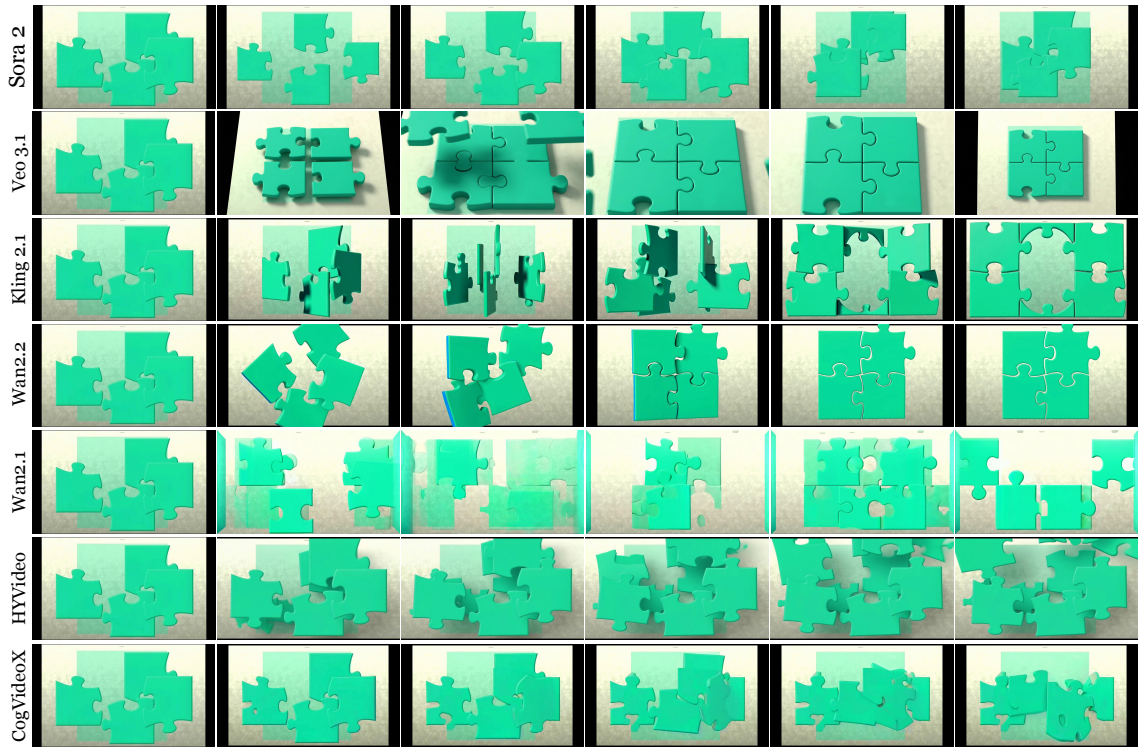


SS-Rule Extrapolation: "The video displays three sets of graphic sequences, each following a specific filling pattern from left to right. One sequence remains incomplete. As time progresses, the question marks in the third sequence gradually disappear, revealing the final graphics that emerge according to the pattern's rules."

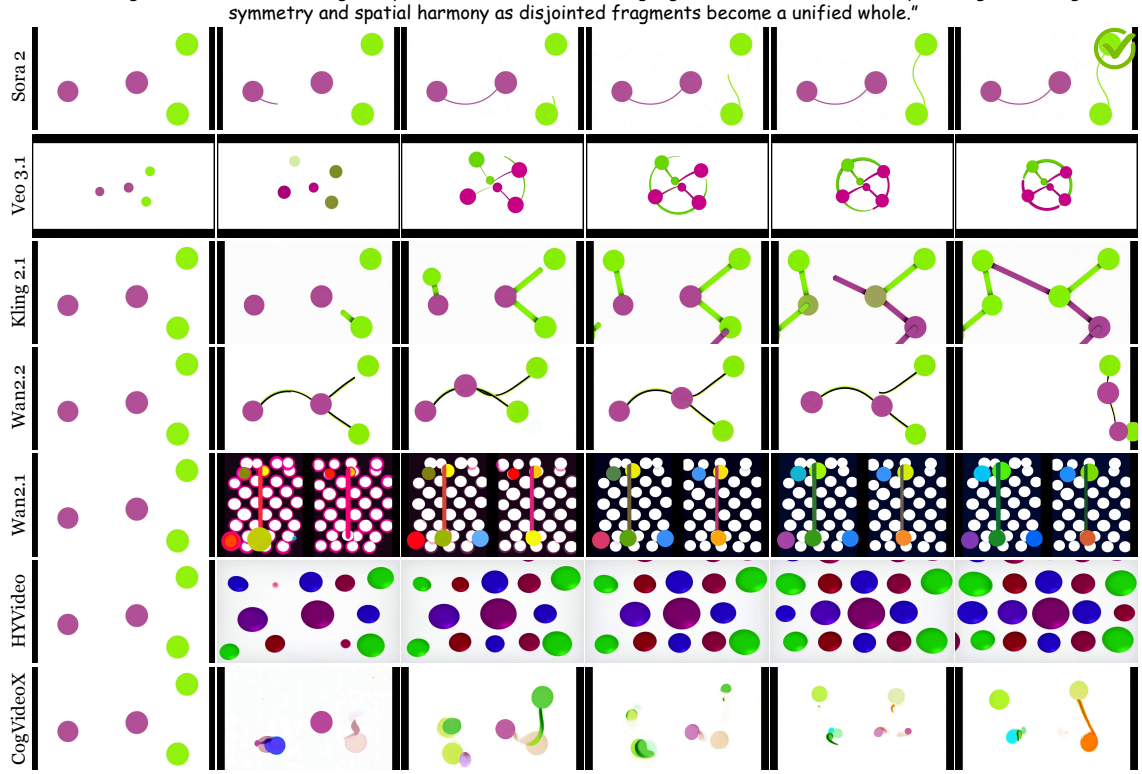


SS-Game Move: "The video shows gameplay footage of a match-three puzzle game. As time progresses, the red blocks that can be eliminated are gradually clicked away, while the remaining blocks shift positions accordingly."

Figure 21. Evaluation case demonstration of Rule Extrapolation and Game Move of Structural Reasoning & Search.

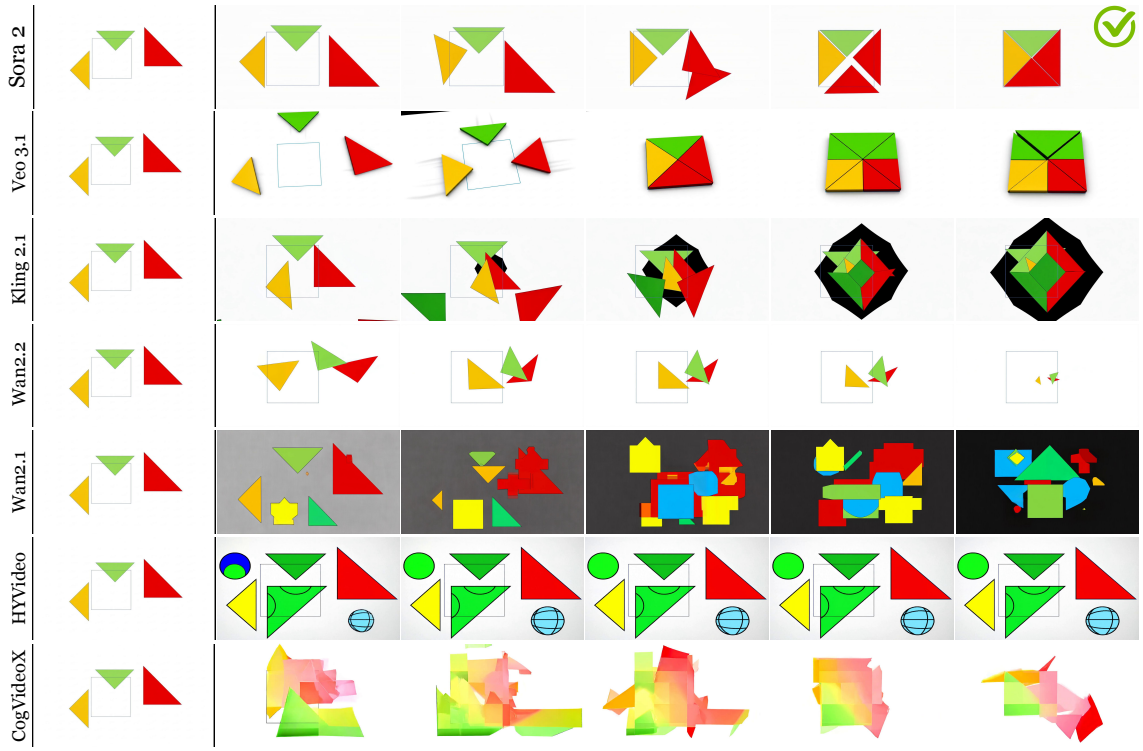


SV-Shape Fitting: "Four vibrant teal puzzle pieces drift and rotate against a textured off-white backdrop, gradually aligning—interlocking tabs and blanks clicking into place. Subtle shadows and highlights trace their motion, emphasizing the emergent symmetry and spatial harmony as disjointed fragments become a unified whole."

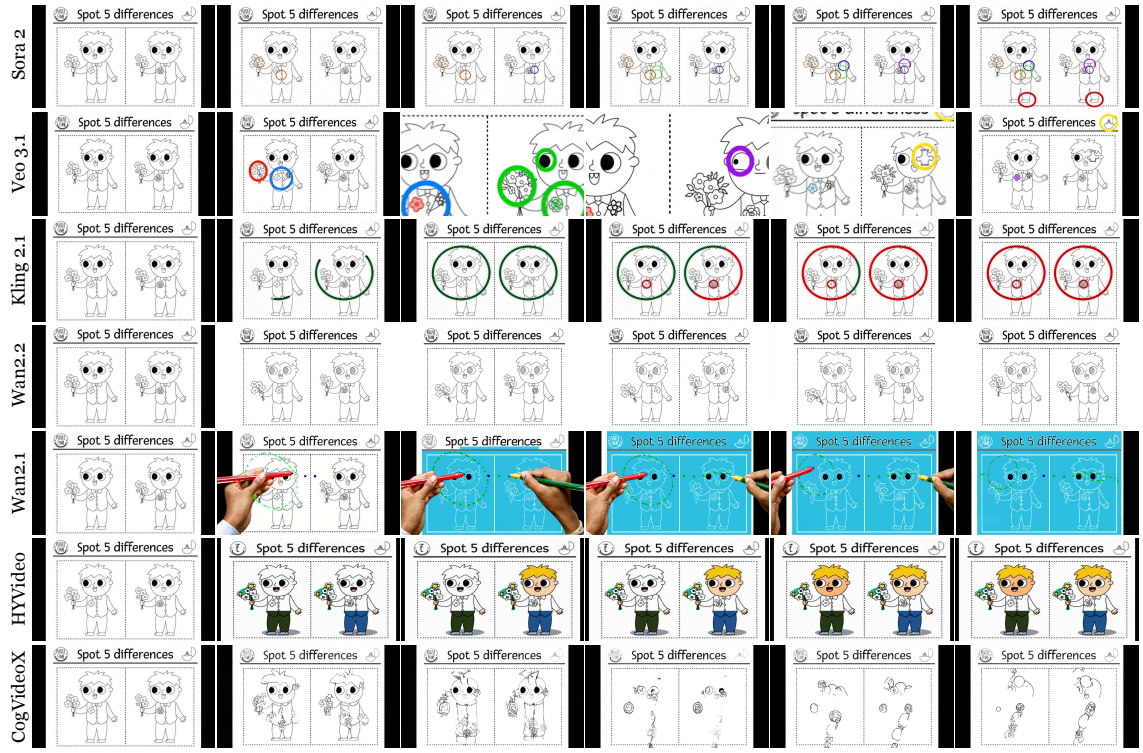


SV-Connecting Colors: "Two curves begin to appear in the video, each connecting two sets of circles of the same color."

Figure 22. Evaluation case demonstration of Shape Fitting and Connecting Colors of Spatial & Visual Pattern Reasoning.

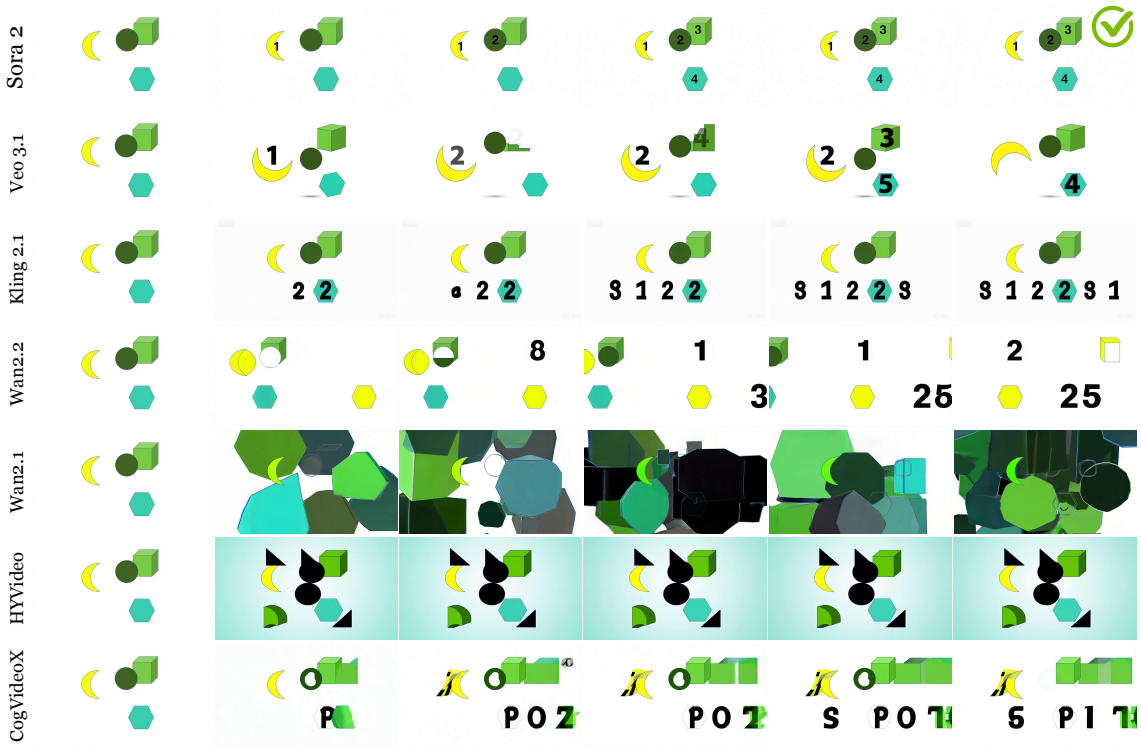


SV-Pattern Recognition: "The colored shapes in the video shift from their scattered positions and eventually align tightly with the black outline in the background."

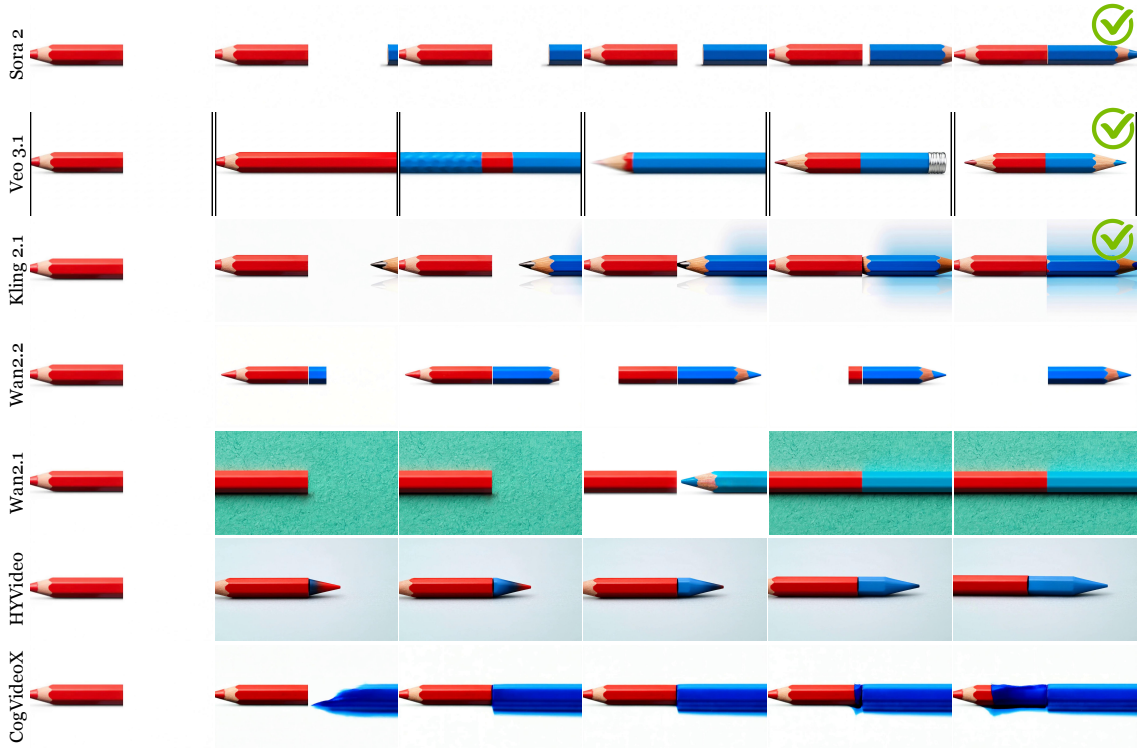


SV-Odd-one-out: "The video displays a spot-the-difference board. As time progresses, the differences on both sides are gradually circled. Corresponding areas are highlighted with circles of the same color."

Figure 23. Evaluation case demonstration of Pattern Recognition and Out-one-out of Spatial & Visual Pattern Reasoning.

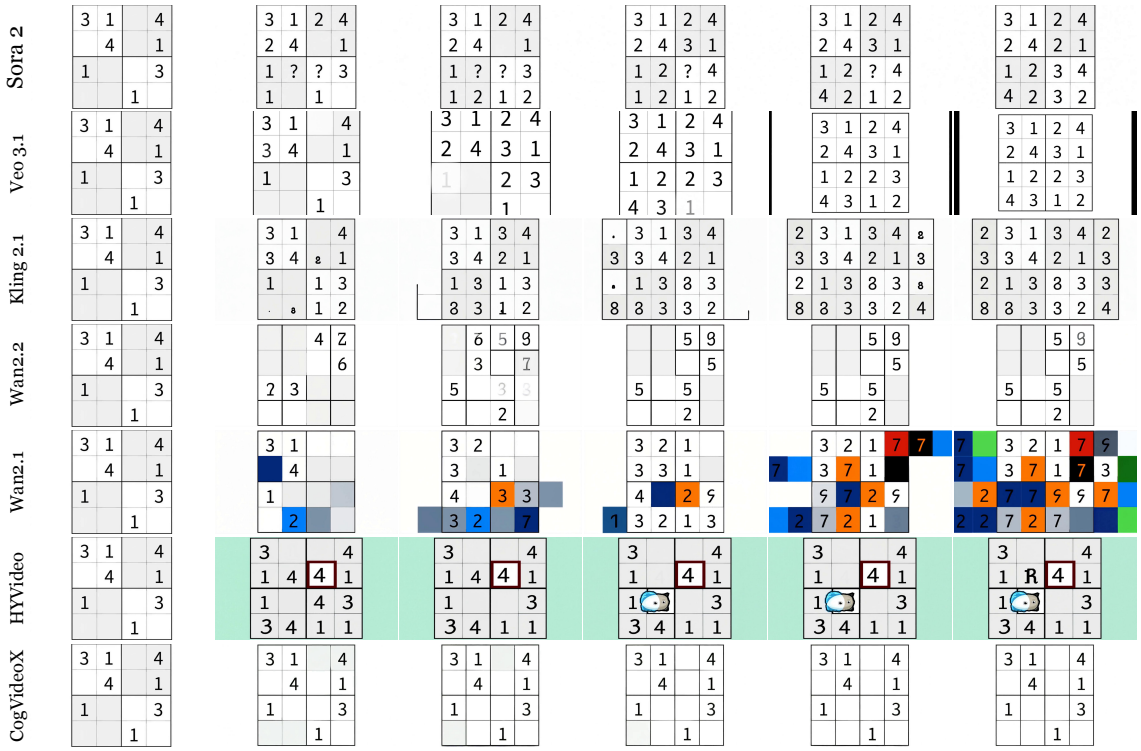


SV-Counting Objects: "The video displays multiple shapes, with black numbers appearing sequentially over time, ranging from 1 to the total number of shapes in the scene."

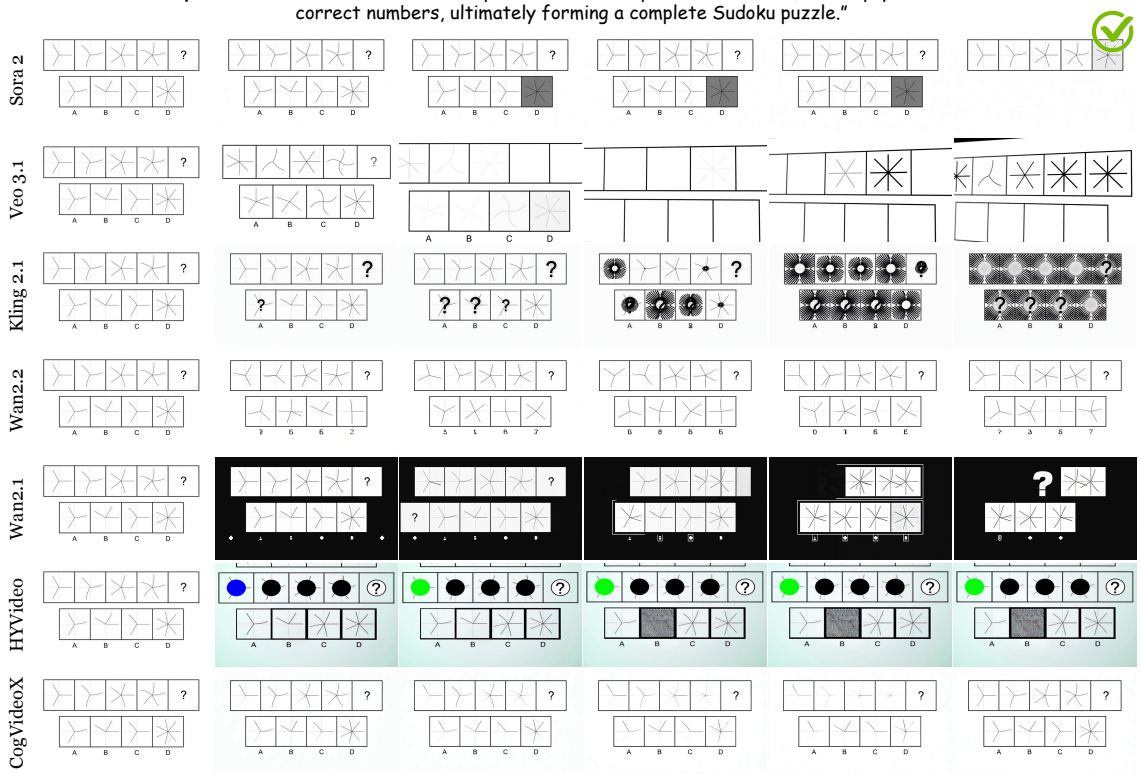


SV-Visual Analogy: "A red pencil lies horizontally, its tip pointing left. Gradually, a blue segment emerges from the right, seamlessly fusing with the red half—splitting the pencil into two vivid, equal halves. The wood grain and sharpened tips remain intact, creating a symmetrical, color-blocked form. The transition feels organic, as if the pencil is being painted or assembled mid-frame, inviting viewers to imagine how the halves connect or rotate to form a whole."

Figure 24. Evaluation case demonstration of Counting Objects and Visual Analogy of Spatial & Visual Pattern Reasoning.

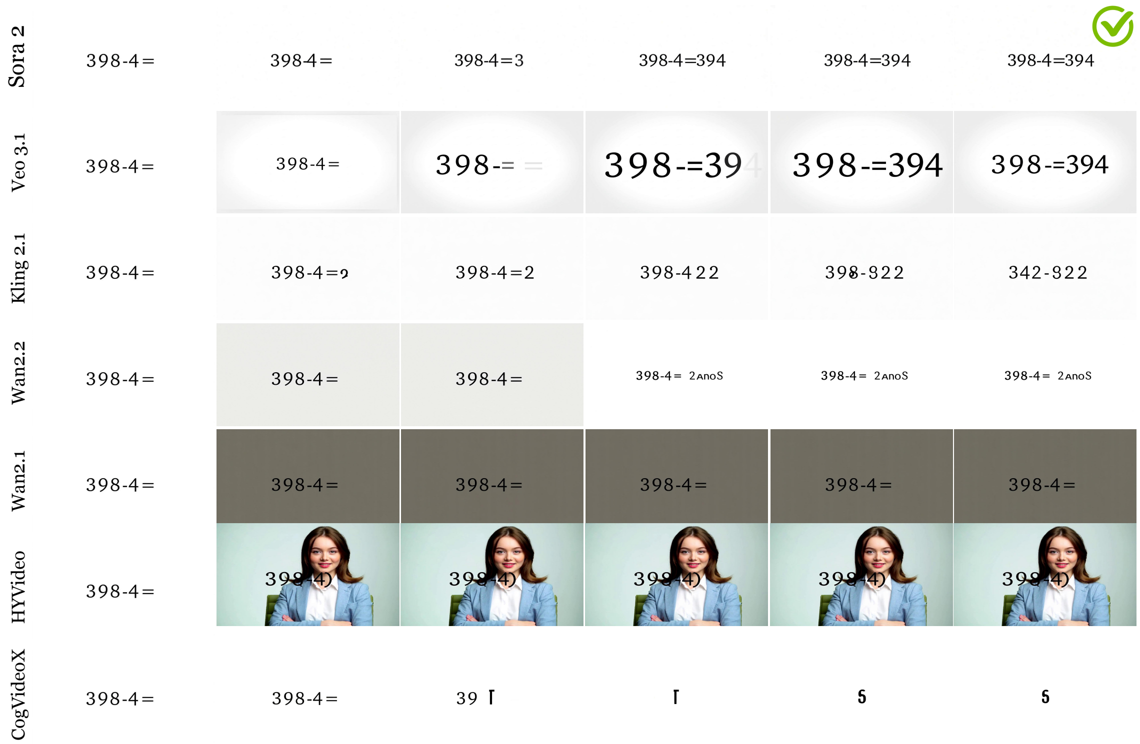


SL-Sudoku Completion: "The video shows an incomplete 4x4 Sudoku puzzle. Over time, the empty cells are filled with the correct numbers, ultimately forming a complete Sudoku puzzle."

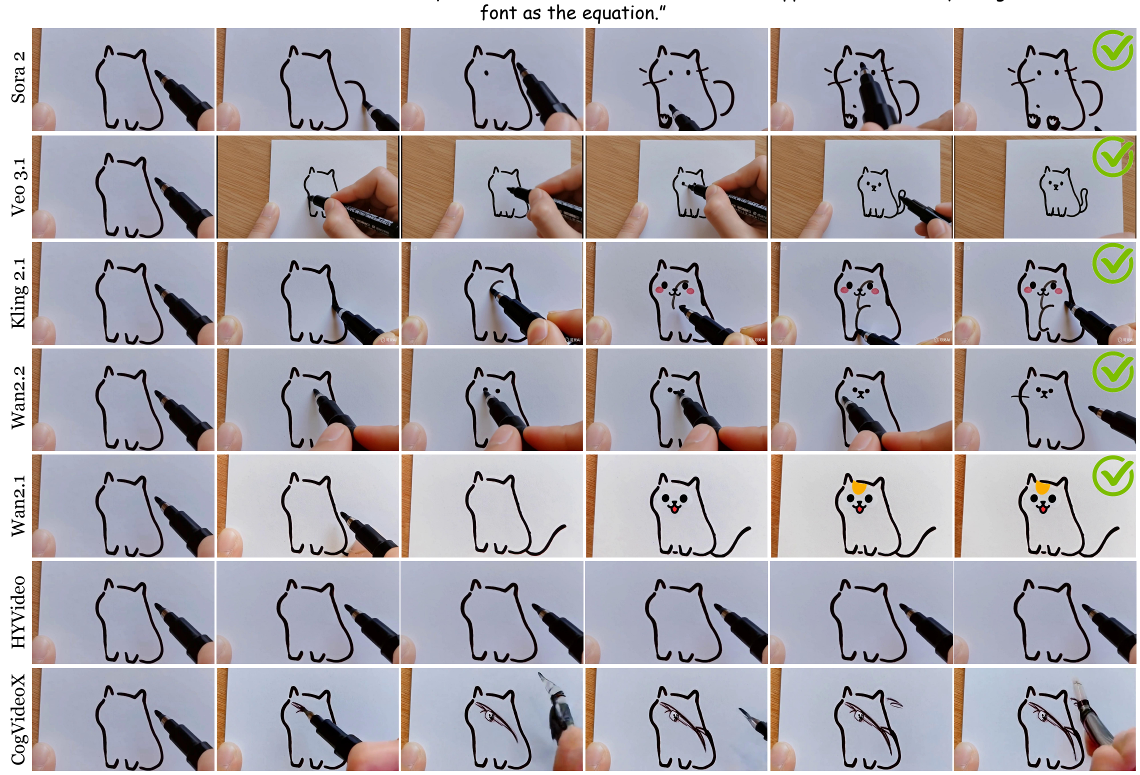


SL-Symbolic Reasoning: "The video shows a sequence of blocks, one of which is marked with a question mark. The options in this block are arranged below the entire sequence. Over time, the question mark and the four options below it gradually disappear, and at the same time, a pattern from the options that forms a regular pattern with the previous (or next) sequence appears in the block where the question mark originally appeared."

Figure 25. Evaluation case demonstration of Sudoku Completion and Symbolic Reasoning of Symbolic & Logical Reasoning.



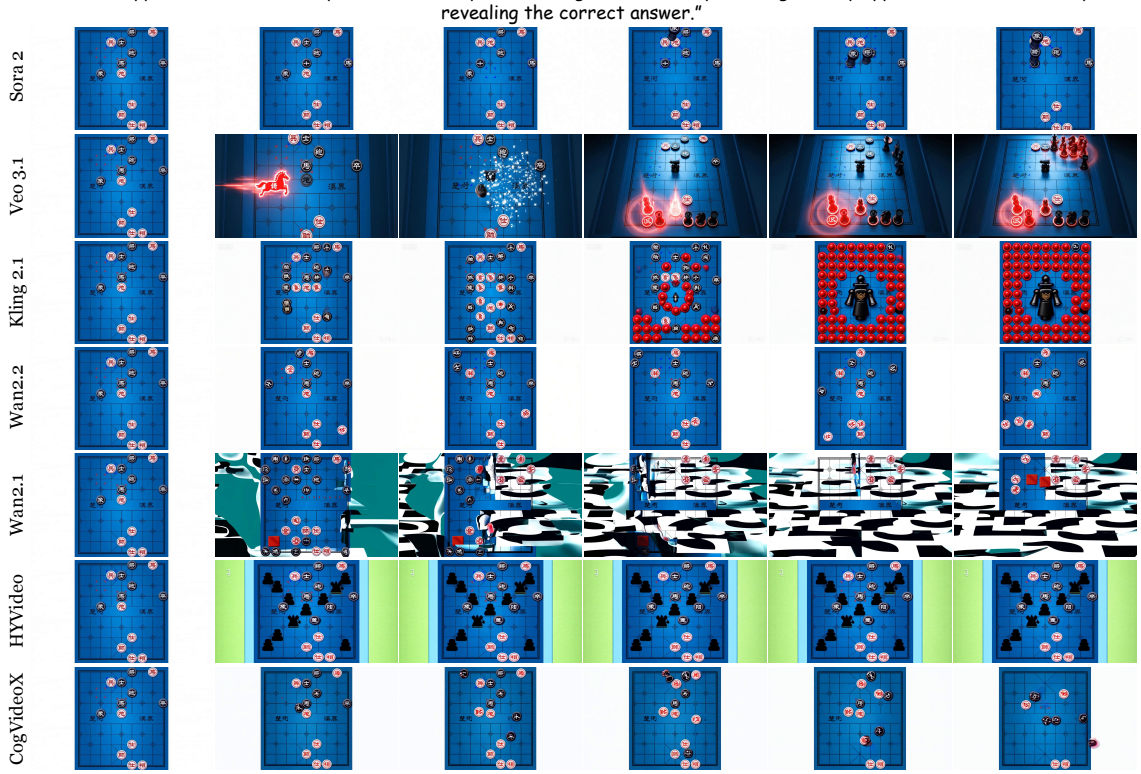
SL-Arithmetic: "The video shows an unsolved equation. Over time, the correct answer appears after the equal sign, in the same font as the equation."



SL-Visual Deduction: "The video shows an unfinished sketch. As time goes by, the sketch is gradually completed, and finally forms a cute cat sketch."

Figure 26. Evaluation case demonstration of Arithmetic and Visual Deduction of Symbolic & Logical Reasoning.

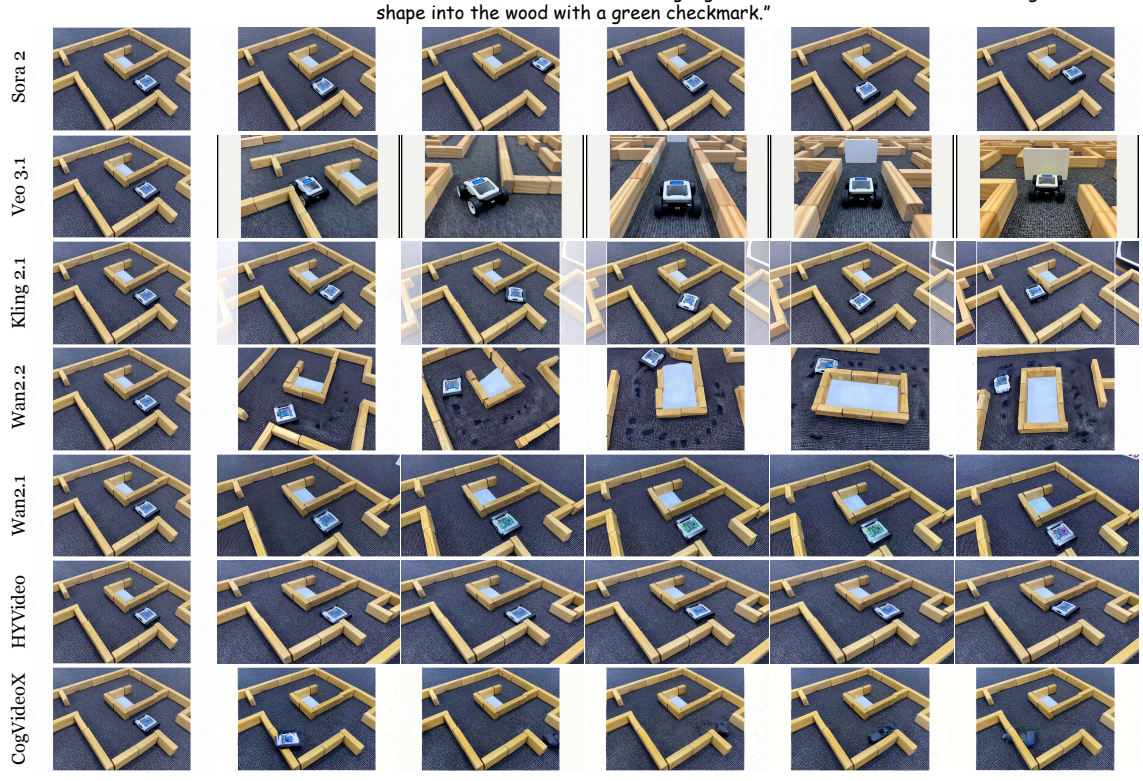
SL-Transtive Reasoning: "The video displays two math problems, one solved and one unsolved. As time progresses, following the solution approach of the solved problem, the steps for solving the unsolved problem gradually appear below it, ultimately revealing the correct answer."



SL-Game Rule: "A strategic chess-like game unfolds: red pieces maneuver under silent constraints, shifting positions to corner the black general. Observe how each move—subtle, precise—alters the board's balance. Symbols vanish, reappear, or lock into place as logic dictates. Trace the invisible rules guiding red's path to victory."

Figure 27. Evaluation case demonstration of Transitive Reasoning and Game Rule of Symbolic & Logical Reasoning.

AT-Tool Use: "The video demonstrates various tools, and over time, it highlights the most suitable tool for cutting a heart shape into the wood with a green checkmark."



AT-Robot Navigation: "The small car follows the maze to reach the white finish point."

Figure 28. Evaluation case demonstration of Tool Use and Robot Navigation of Action Planning & Task Execution.

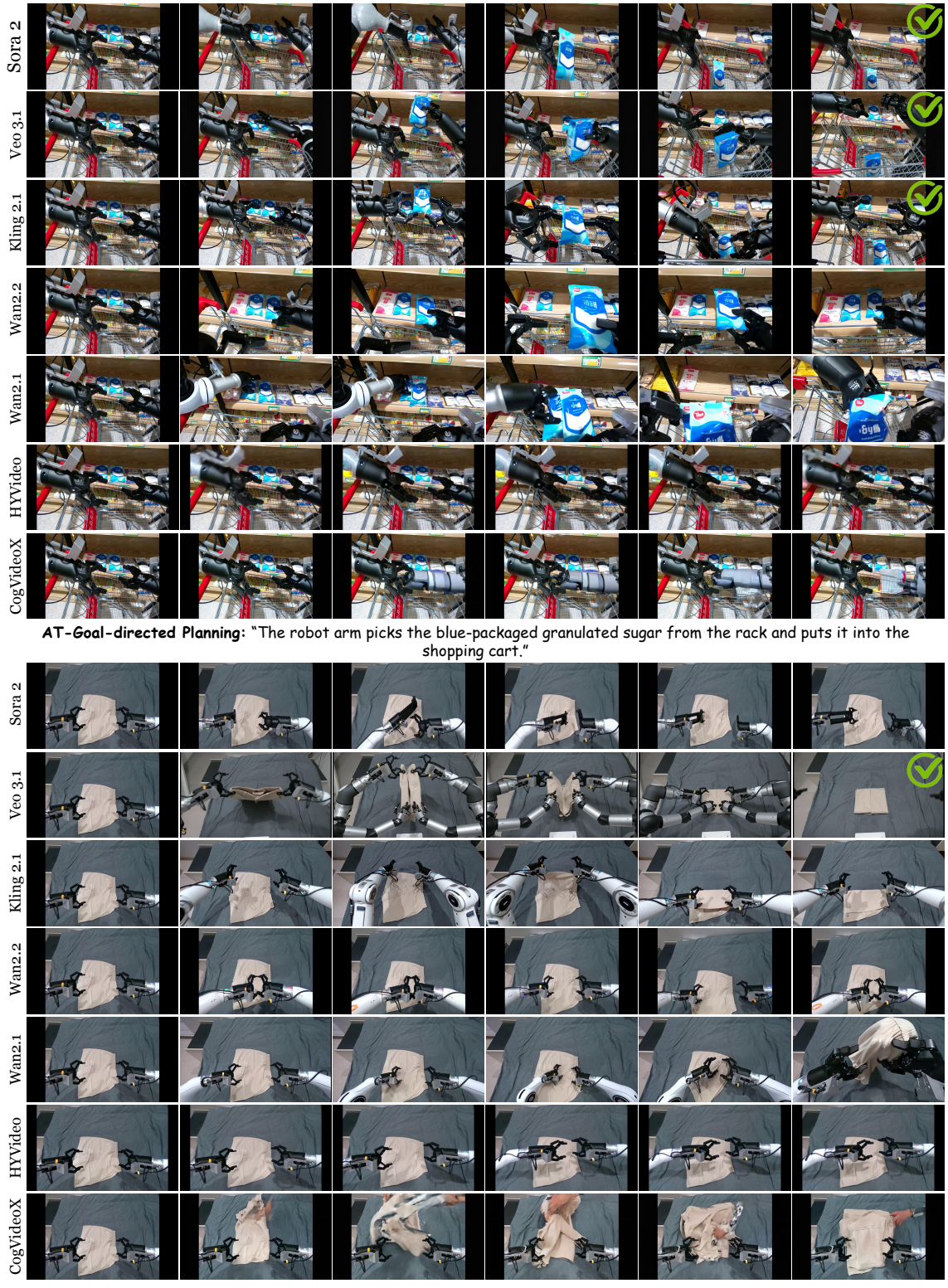
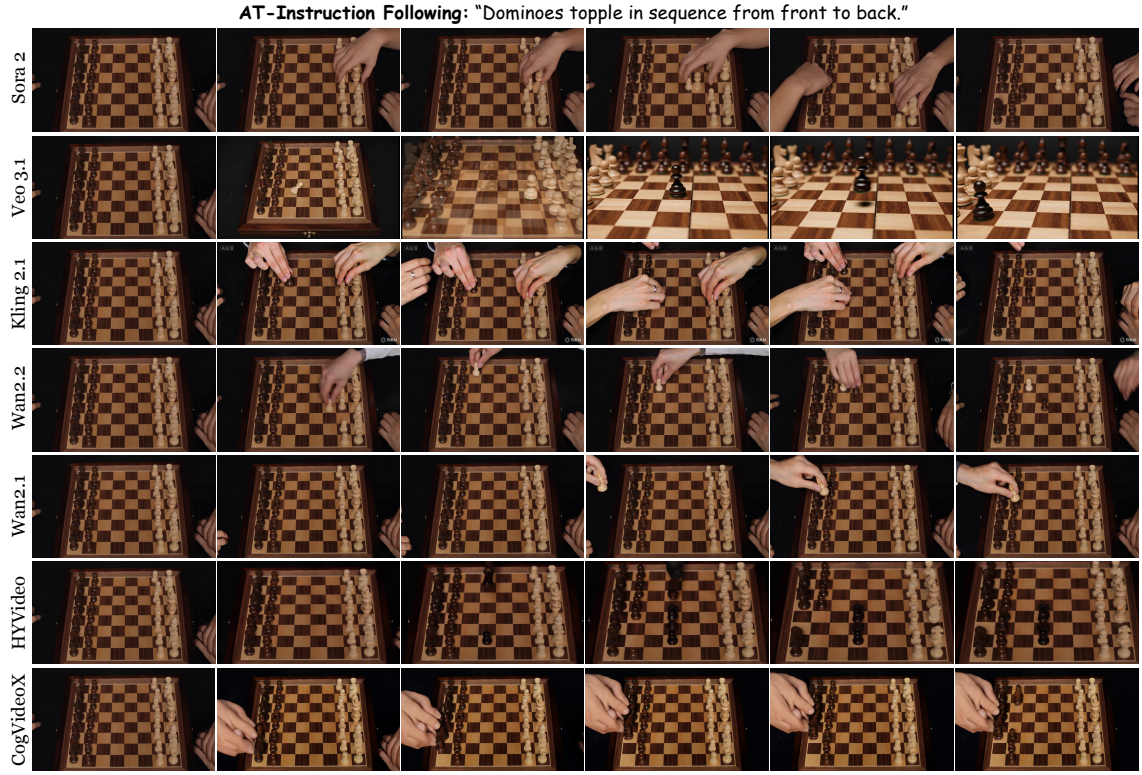


Figure 29. Evaluation case demonstration of Goal-directed Planning and Multi-step Manipulation of Action Planning & Task Execution.

AT-Instruction Following: "Dominoes topple in sequence from front to back."



AT-Game Strategy: "White moves the pawn in front of the king forward one square; Black moves the pawn in front of the bishop two squares."

Figure 30. Evaluation case demonstration of Instruction Following and Game Strategy of Action Planning & Task Execution.