Thinking with Video: Video Generation as a Promising Multimodal Reasoning Paradigm

Supplementary Material

A. Related Work

Video Generation Model: The field of video generation is advancing rapidly. Early models like OpenAI's Sora are the "GPT-1 moment [24]" for video, and now newer versions like Sora-2 have made a huge leap forward. Sora-2 can create more realistic and controllable videos that are physically accurate and even include synchronized dialogue and sound effects. Besides Sora, other powerful but closed-source models are pushing the industry forward. Companies like Runway, with its Gen-3 model [26], Pika Labs, Luma AI, and Google DeepMind's Veo [10] series are all creating impressive, high-quality videos. However, because these models are proprietary, they are not widely available for researchers to study and build upon. To counter this, a movement of open-source alternatives is growing. Projects like Stable Video Diffusion [2], Hunyan-Video [15], and the Wan series [31] are making video generation technology accessible to everyone.

Reasoning Paradigm Transfer: Chain-of-Thought (CoT) significantly improves the reasoning ability of large language models (LLMs) [12, 33, 35, 40, 43]. Large-scale reinforcement learning incentivizes LLMs to think productively using their CoT [12, 22, 46]. o3 and o4-mini further extend this capability by natively "Thinking with Images" in their CoT, which involves directly cropping, zooming, and rotating images [23]. "Thinking with Images [16, 20, 23, 45]" is a paradigm that outputs images in CoT to help VLMs reason better, largely improving the VLMs' reasoning abilities [30]. Recently, unified multimodal understanding and generating models have appeared [6, 7, 38, 42, 44]. They potentially achieve "Thinking with Images" through text and image interleaved reasoning.

Evaluation of Video Generation Reasoning: Video-generation-based reasoning has only recently begun to be explored [13, 36]. Wiedemer et al. show that Veo 3 can solve many tasks it was not specifically trained for. These abilities span perceiving, modeling, and manipulating the visual world, enabling early forms of video-based reasoning. Their evaluations include tasks such as maze solving and visual symmetry.

However, existing works [13, 36] differ from our focus in several key aspects: (1) Vision-centric scope: Their evaluations primarily focus on vision-centric reasoning tasks and do not extend to text-centric or broader multimodal reasoning settings. (2) Case-based evaluation setup: These works include both qualitative demonstrations and several quantitative evaluations. However, the evaluations are conducted on a limited number of manually curated scenes or canonical examples with restricted diversity. As a result, each task is tested on relatively small sample sizes, making it challenging to assess generalization or statistical robustness. (3) Lack of systematic comparison with VLMs: These works do not provide a systematic comparison with SOTA Vision-Language Models (VLMs) across diverse task categories, leaving the relative strengths of video vs. vision-language models underexplored.

Our work complements these directions with the following contributions: (1) Unified multimodal reasoning paradigm: We evaluate video models not only on vision-centric tasks but also on text-centric and multimodal reasoning tasks, demonstrating that video generation may serve as a general multimodal reasoning paradigm rather than a purely visual one. (2) Systematic and verifiable benchmark construction: We systematically construct the VideoThinkBench where large numbers of test cases can be generated in batches using a program. Most of the vision-centric tasks we have designed are verifiable. (3) Systematic comparison with VLMs: We conduct comprehensive comparisons with SOTA VLMs, providing the first systematic study of how "Thinking with Video" behaves relative to "Thinking with Images". In summary, we propose "Thinking with Video" as a new paradigm with the potential to unify multimodal reasoning. Furthermore, we find that video model reasoning can be enhanced through few-shot learning and test time scaling (self-consistency).

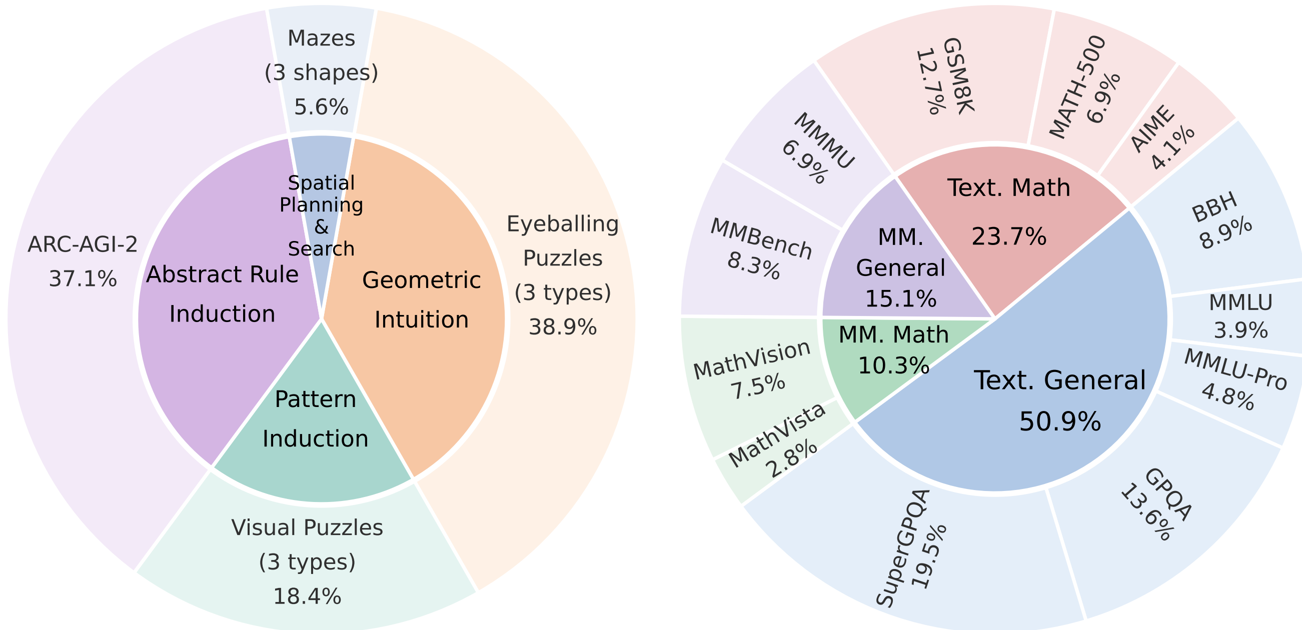
B. Limitations and Future Work

We primarily evaluate Sora-2's reasoning abilities among video generation models. Sora-2 is not open-source, limiting the analysis of its internal mechanisms.

For future evaluation work, we plan to include more video generation models, especially open-source models. This allows for a deeper analysis of their internal mechanisms. Meanwhile, there are other capabilities of video models worth exploring.

To enhance the reasoning abilities of video models through training, a promising direction is to scale up the verifiable tasks in VideoThinkBench via Reinforcement Learning with Verifiable Rewards (RLVR), thereby enhancing models' "Thinking with Video" capabilities.

Regarding unified multimodal training for video models, we will explore converting textual corpora into video-form training data (e.g., by generating the next word frame-by-



(a) **Vision-centric reasoning tasks**

(b) **Text-centric reasoning tasks**

Figure 8. Task composition and distribution of Video Thinking Benchmark (VideoThinkBench). (a) Vision-centric tasks contain tasks that we design (e.g., Eyeballing Puzzles) and tasks adapted from existing benchmarks (e.g., ARC-AGI-2), evaluating four abilities. (b) Text-centric tasks consist of subsets sampled from text-only and multimodal reasoning benchmarks, adapted for video generation reasoning. The former contains math reasoning (Text. Math) and general knowledge reasoning (Text. General) benchmarks and the latter also contains math reasoning (MM. Math) and general knowledge reasoning (MM. General) benchmarks.

frame to simulate whiteboard handwriting). The idea is that by pretraining video generation models on such text-generation tasks, they can acquire textual world knowledge. Ultimately, with large-scale image-text data training, these models might achieve unified multimodal understanding and generation.

C. More about VideoThinkBench

C.1. Detailed Sample Distribution

VideoThinkBench contains 4,149 test samples in total. Vision-centric tasks contain 2,696 samples and text-centric tasks contain 1,453 samples in total. For text-centric tasks, we sampled a subset from most of the selected benchmarks for evaluation cost control. Task distribution is illustrated in Figures 8a and 8b, with detailed statistics listed below.

Vision-Centric Tasks: Eyeballing Puzzles (1,050), Visual Puzzles (496), ARC-AGI-2 (1,000), Mazes (150).

Text-Centric Tasks

- Text-Only Math Reasoning (345 samples): GSM8K (185) [3]; MATH-500 (100) [4]; AIME24 (30); AIME25 (30).
- Text-Only General Knowledge Reasoning (739 samples): BBH (130) [28]; MMLU (57) [14]; MMLU-Pro (70) [34]; GPQA-diamond (198) [25]; SuperGPQA-easy (284) [8].
- Multimodal Reasoning (369 samples): MathVista (40) [18]; MathVision (109) [32]; MMBench (120) [17]; MMMU (100) [41].

C.2. Mini Test Set

We construct a mini test set to reduce the evaluation cost, making our benchmark easier for researchers to use. This mini test set is a subset of the full set and covers all the benchmark tasks, with 750 test samples in total:

Vision-Centric Tasks (500 samples): Eyeballing Puzzles (210, 10 per task), Visual Puzzles (100, 10 per task), ARC-AGI-2 (140), Mazes (50, covering three maze shapes).

Text-Centric Tasks: 250 samples of the full set.

D. More Evaluation Results

We test more models on the mini test set of VideoThinkBench (Sec. C.2), including more video generation models (Seedance 1.0 Pro [9], MiniMax Hailuo 2.3 [19], Wan2.2-TI2V-5B [31] and MOVA [29]), the image generation models (Nano Banana Pro [11], Seedream 4.5 [27], GPT Image 1.5 [21], BAGEL [7] and Qwen-Image-Edit-2511 [37]), and Qwen3-VL series [1]. The results are shown in Tab. 7 and Tab. 8.

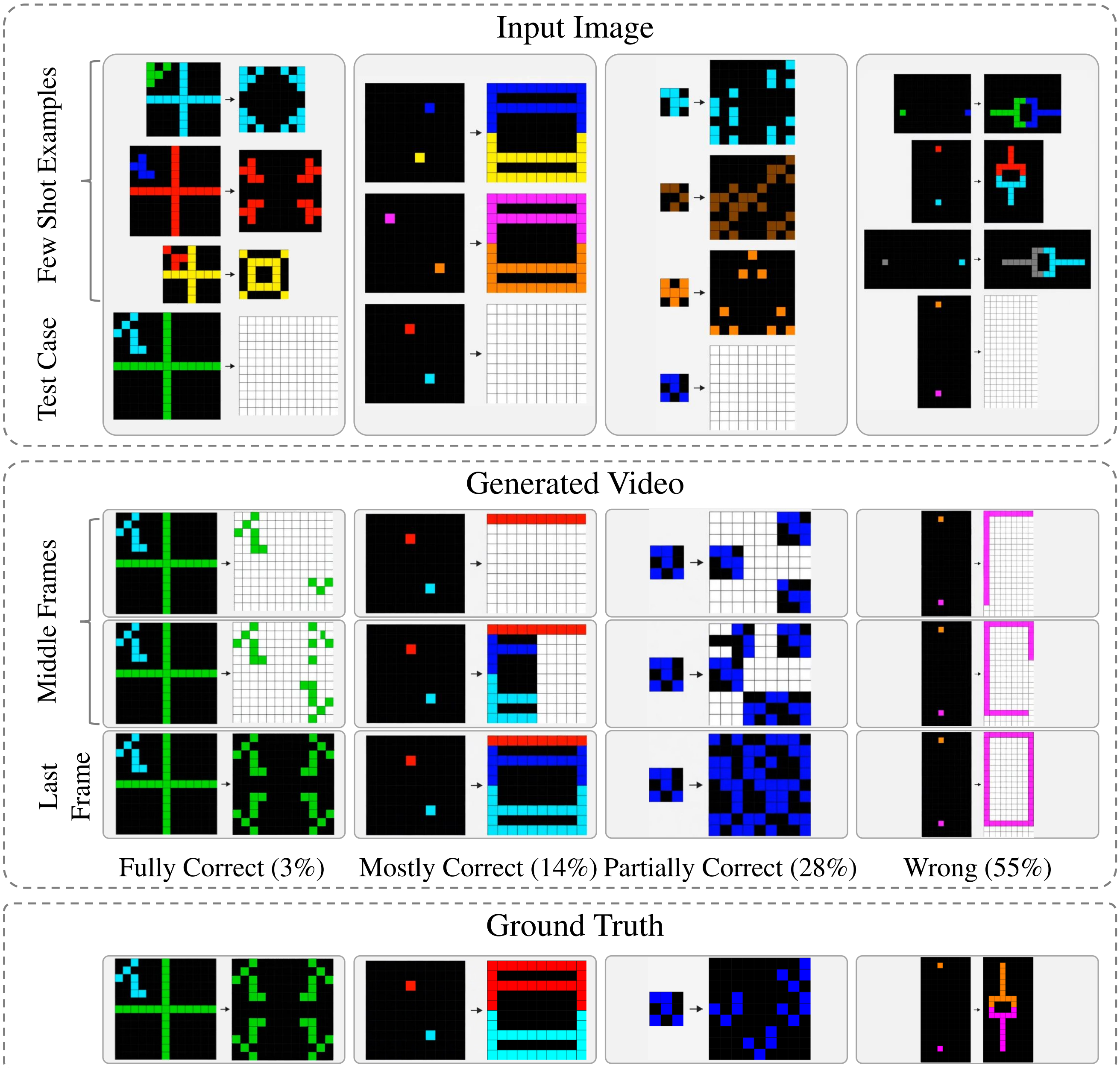


Figure 9. **Examples of Sora-2 trying to solve ARC-AGI-2.** ARC-AGI-2 is a benchmark targeting few-shot, inductive reasoning over abstract pattern transformations. Sora-2 is expected to deduce the transformation rule from the examples and use it to generate the output grid for the test case. When manually evaluating, we classify the correctness of the answer from Sora-2 into 4 categories. Details: Sec. 2.1.3

E. Detailed Evaluation Protocols

This section provides comprehensive details on the evaluation protocols for all tasks in VideoThinkBench. We present the dataset construction methods, evaluation procedures, and prompts used for both video generation models and VLM baselines. The tasks are organized by category: spatial reasoning tasks (Mazes and Eyeballing Puzzles), inductive reasoning tasks (Visual Puzzles and ARC-AGI-2), and text-centric reasoning tasks.

E.1. Generation Parameters

For Sora-2, the video duration is 10 seconds in all the experiments. For evaluation of Wan 2.5, as detailed in Sec. 3.2.3, we use the model of wan2.5-i2v-preview, setting the resolution to 480P and the duration to five seconds.

Table 7. Accuracy (%) of different models across the vision-centric tasks of VideoThinkBench (mini test set).

Model	Average	Eyeballing Puzzles			Visual Puzzles			ARC-AGI-2	Mazes		
		Point	Line	Shape	Symmetry	Gradient	Comp.		Square	Hexagon	Circle
Video Generation Models											
Sora-2	**31.6**	50.0	35.0	25.0	80.0	35.0	53.0	2.8	35.3	0.0	0.0
Veo 3.1	**27.7**	34.4	24.3	30.0	77.5	40.0	70.0	0.7	0.0	0.0	0.0
MiniMax Hailuo 2.3	**25.9**	36.7	34.3	27.5	72.5	45.0	42.5	0.0	0.0	0.0	0.0
MOVA-360p	**13.4**	23.3	25.7	25.0	45.0	0.0	15.0	0.0	0.0	0.0	0.0
Seedance 1.0 Pro	**12.4**	22.2	24.3	35.0	25.0	10.0	7.5	0.0	0.0	0.0	0.0
MOVA-720p	**11.8**	32.2	18.6	25.0	30.0	0.0	12.5	0.0	0.0	0.0	0.0
Wan2.2-TI2V-5B	**7.3**	17.8	10.0	20.0	7.5	10.0	7.5	0.7	0.0	0.0	0.0
Image Generation Models											
Nano Banana Pro	**29.8**	24.0	30.0	35.0	85.0	50.0	73.0	0.7	0.0	0.0	0.0
Seedream 4.5	**24.4**	25.6	16.3	30.0	75.0	35.0	62.5	0.0	0.0	0.0	0.0
GPT Image 1.5	**19.2**	24.4	15.0	17.5	38.0	50.0	47.5	0.0	0.0	0.0	0.0
Qwen-Image-Edit-2511	**14.9**	30.0	23.8	27.5	25.0	35.0	7.5	0.0	0.0	0.0	0.0
BAGEL (Image Output)	**7.7**	24.4	12.5	25.0	5.0	0.0	10.0	0.0	0.0	0.0	0.0
Vision-Language Models											
Claude Sonnet 4.5	**37.3**	40.0	34.0	60.0	75.0	75.0	83.0	5.7	0.0	0.0	0.0
Gemini 2.5 Pro	**35.6**	33.0	23.0	40.0	95.0	95.0	68.0	2.1	0.0	0.0	0.0
GPT-5 high	**35.5**	39.0	30.0	23.0	98.0	80.0	85.0	0.0	0.0	0.0	0.0
Qwen3-VL-235B-A22B	**30.2**	24.0	17.0	30.0	93.0	55.0	83.0	0.0	0.0	0.0	0.0
Qwen3-VL-32B	**29.6**	33.0	21.0	20.0	85.0	55.0	78.0	4.1	0.0	0.0	0.0
Qwen3-VL-Plus	**29.4**	32.0	29.0	30.0	90.0	35.0	78.0	0.0	0.0	0.0	0.0

E.2. Spatial Reasoning Tasks

E.2.1. ARC-AGI-2

We present four cases and the corresponding manual evaluation category in Fig. 9.

E.2.2. Mazes

Dataset Construction We use programs to automatically construct a dataset of 150 mazes, divided equally into three distinct geometric types: square mazes, hexagon mazes, and circle mazes. For each type, we generated 50 unique instances, each with a start and end point marked by red dots. The task requires the model to generate a path from start to end while not overlapping black walls.

Evaluation Setup Evaluation is conducted automatically on the final frame of the generated video. A solution is considered successful only if it satisfies two conditions: 1) Red pixels form a continuous line connecting the start and end points. 2) No red pixel overlaps any black pixel in the input image which represents the maze walls. An attempt is marked correct only if both criteria are fully met.

Evaluation Results Sora-2's performance on the maze-solving task varied significantly depending on the maze's geometric structure. As shown in Fig. 10, it demonstrated a moderate ability to solve traditional square mazes, successfully finding a valid path in 20 out of 50 instances for a

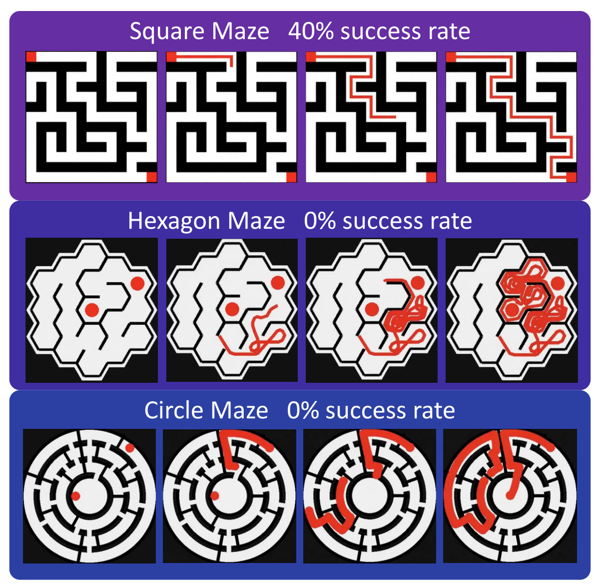


Figure 10. **Examples and evaluation results of Mazes.** In each quartet, the first image is the input image, and other three images are from videos generated by Sora-2. We generate and evaluate 50 samples for each type of maze. Prompt: *"Draw a red path connecting two red dots without touching the black walls. In portrait. Static camera."* Sora-2 successfully solves the square maze but fails at other two mazes. Details: Sec. E.2.2.

Table 8. Accuracy (%) of different models across the text-centric tasks of VideoThinkBench (mini test set).

Model	Average	Text-Only Math Reasoning				Text-Only General Reasoning					MM. Math Reason.		MM. General Reason.	
		GSM8K	MATH-500	AIME24	AIME25	BBH	MMLU	MMLU-Pro	GPQA	SuperGPQA	MathVista	MathVision	MMBench	MMMU
Video Generation Models														
Sora-2 (Audio)	**67.8**	98.9	92.0	46.7	36.7	80.6	67.3	76.5	57.6	44.5	75.7	46.7	89.0	69.2
Sora-2 (Last Frame)	**57.0**	75.7	67.0	38.3	33.3	69.8	69.1	72.0	51.5	53.2	67.6	44.9	60.4	38.3
Veo 3.1 (Last Frame)	**48.3**	80.0	70.0	50.0	20.0	61.5	16.7	60.0	52.0	42.9	50.0	45.0	35.0	45.0
Veo 3.1 (Audio)	**44.5**	93.3	80.0	50.0	20.0	61.5	41.7	80.0	40.0	51.4	25.0	5.0	20.0	10.0
MiniMax Hailuo 2.3	**38.4**	76.6	40.0	10.0	20.0	61.5	33.3	86.6	16.0	65.7	30.0	10.0	30.0	20.0
MOVA-720p (Last Frame)	**12.5**	30.0	35.0	20.0	0.0	15.4	0.0	6.7	8.0	2.9	0.0	25.0	10.0	10.0
MOVA-360p (Last Frame)	**10.4**	20.0	10.0	0.0	20.0	15.4	0.0	0.0	20.0	0.0	5.0	30.0	5.0	10.0
MOVA-720p (Audio)	**7.2**	0.0	0.0	0.0	0.0	30.8	16.7	0.0	8.0	2.9	0.0	10.0	20.0	5.0
MOVA-360p (Audio)	**5.8**	0.0	0.0	0.0	0.0	30.8	33.3	0.0	0.0	5.7	0.0	0.0	5.0	0.0
Seedance 1.0 Pro	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Wan2.2-TI2V-5B	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Image Generation Models														
Nano Banana Pro	**66.0**	56.7	65.0	80.0	80.0	69.2	75.0	80.0	44.0	62.9	75.0	45.0	75.0	50.0
Seedream 4.5	**55.7**	100.0	80.0	20.0	10.0	69.2	75.0	60.0	36.0	48.6	55.0	60.0	55.0	55.0
GPT Image 1.5	**41.4**	90.0	40.0	0.0	0.0	69.2	25.0	46.7	40.0	22.9	50.0	40.0	65.0	50.0
Qwen-Image-Edit-2511	**10.5**	0.0	0.0	0.0	0.0	15.4	16.7	6.7	4.0	8.6	15.0	15.0	30.0	25.0
BAGEL (Image Output)	**8.9**	6.7	0.0	0.0	0.0	0.0	0.0	6.7	4.0	2.9	25.0	40.0	10.0	20.0
Vision-Language Models														
Gemini 2.5 Pro	**89.0**	100.0	100.0	100.0	90.0	100.0	83.3	93.3	80.0	80.0	85.0	65.0	95.0	85.0
GPT-5 high	**86.6**	100.0	100.0	100.0	100.0	100.0	83.3	93.3	80.0	83.9	75.0	55.0	85.0	70.0
Qwen3-VL-235B-A22B	**77.6**	100.0	100.0	80.0	50.0	84.6	58.3	100.0	56.0	80.0	70.0	65.0	90.0	75.0
Claude Sonnet 4.5	**77.2**	100.0	100.0	60.0	40.0	100.0	83.3	100.0	60.0	80.0	80.0	45.0	80.0	75.0
Qwen3-VL-Plus	**75.8**	100.0	95.0	100.0	70.0	76.9	66.7	80.0	64.0	57.1	65.0	65.0	80.0	65.0
Qwen3-VL-32B	**72.5**	100.0	95.0	80.0	50.0	76.9	66.7	93.3	40.0	65.7	75.0	45.0	90.0	65.0

40% success rate. However, the model's spatial reasoning did not extend to other geometries. For both the hexagon and circle mazes, Sora-2 failed to produce a single correct solution, resulting in a **0% success rate** for both categories. This stark performance gap suggests that while Sora-2 can handle basic pathfinding on grid-like structures, its reasoning struggles to adapt to more complex shapes.

E.2.3. Eyeballing Puzzles

Overview We show examples of all 21 eyeballing puzzle types in Fig. 11.

Evaluation Details For Sora-2, the three evaluation methods on Eyeballing Puzzles are introduced as follows:

- **Audio Evaluation:** The prompt instructs the model to speak the option using the phonetic alphabet ("Alpha", "Bravo", "Charlie", "Delta" and "Echo"). The audio is extracted from the generated video and transcribed using the whisper-1 model. A program then finds the first phonetic-alphabet token in the transcript and treats it as the audio answer. Finally, the extracted answer is compared with the ground truth.
- **Last Frame Evaluation:** The prompt instructs model to draw a red dot on correct option. The last frame of generated video is fed to an image evaluator program that calculates average coordination of red pixels. The last frame option is the option nearest to average coordination of red pixels, or none if there are no red pixels found. Finally, Compare the last frame option with ground truth.
- **Major Frame Evaluation:** For every 5 frames in the video, one frame is extracted and fed to the image evaluator, getting option of this frame. Major frame option is the majority vote result of all chosen frames. "None" option is excluded from voting. Finally, the major-frame option is compared with the ground truth.

Results in Table We present the results of eyeballing tasks in Tab. 9.

E.3. Inductive Reasoning Tasks

E.3.1. Visual Puzzles

Formal Definitions of the Deviation Value In visual puzzles, the deviation value $Diff$ is defined to quantify the deviation between a generated video frame and the ground truth image. This metric is computed as **the sum of per-pixel differences within the puzzle area**. Formally:

$$Diff = \sum_{(x,y) \in \text{Puzzle Area}} \delta(\text{Pixel}_{\text{gen}}(x,y), \text{Pixel}_{\text{gt}}(x,y)) \tag{1}$$

The per-pixel difference function δ is defined according to the task type:

- For color-filling tasks: We calculate **the Euclidean distance in RGB space**:

$$\delta_{\text{color}}(p,q) = \sqrt{(p_r - q_r)^2 + (p_g - q_g)^2 + (p_b - q_b)^2} \tag{2}$$

  where p and q are the pixels from the generated and ground truth images, respectively.
- For shape-drawing tasks: The images are first converted to grayscale. Then we binarize the images and compute

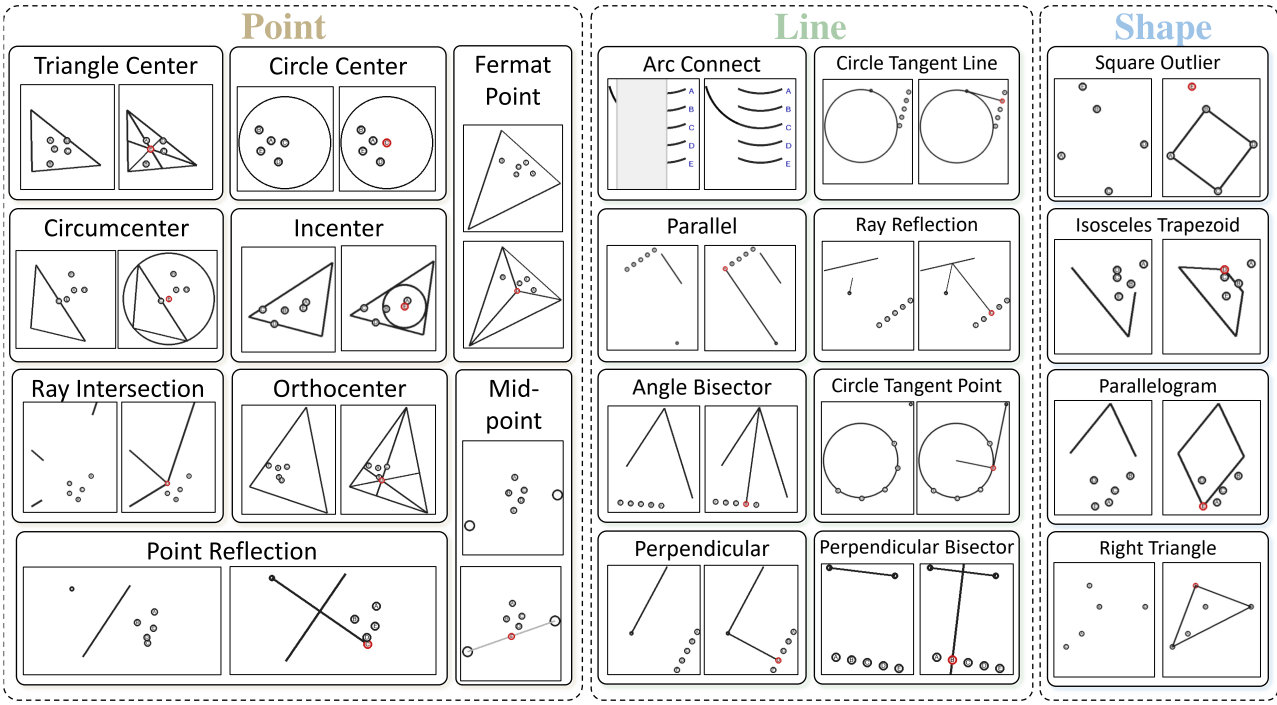


Figure 11. **Overview of 21 eyeballing puzzle types.** Based on the task requirement (constructing a point, a line, or a shape), we divide the puzzle types into Point, Line, and Shape categories. For each puzzle type, an input image and corresponding ground truth image is shown. All prompts: Sec. E.2.3.

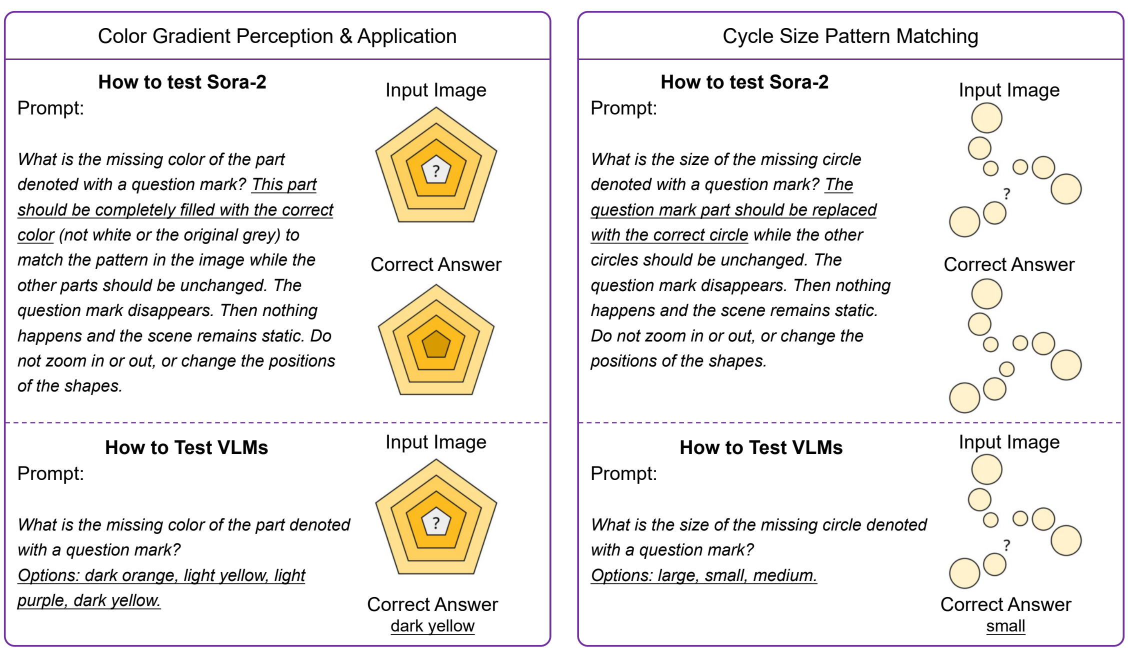


Figure 12. Two examples from the five visual puzzle tasks (listed in Sec. E.3.1) where **tested VLMs are provided with the multiple-choice options**. VLMs only need to select the correct option while Sora-2 needs to correctly solve the task in the generated video.

an **"coverage difference"**, where a pixel is considered different if its binarized color (black/white) differs:

$$\delta_{\text{shape}}(p, q) = \begin{cases} 1, & \text{if} \text{Binarize}(p) \neq \text{Binarize}(q) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Table 9. Accuracy (%) of Sora-2 using 3 evaluation methods and 3 VLMs on eyeballing tasks. The highest score in each row is highlighted in bold. Details: Sec. 2.1.1

Task	Sora-2 Audio	Sora-2 Last Frame	Sora-2 Major Frame	Gemini 2.5 Pro	GPT-5 high	Claude Sonnet 4.5
			Point Tasks			
Ray Intersection	22.0	70.0	**88.0**	22.0	16.0	22.0
Midpoint	22.0	48.0	64.0	28.0	34.0	**66.0**
Circle Center	58.0	56.0	**70.0**	44.0	62.0	50.0
Point Reflection	18.0	22.0	22.0	**30.0**	28.0	**30.0**
Triangle Center	34.0	42.0	**44.0**	38.0	40.0	36.0
Incenter	**48.0**	30.0	34.0	32.0	30.0	34.0
Circumcenter	14.0	20.0	24.0	12.0	**32.0**	26.0
Orthocenter	**32.0**	18.0	26.0	14.0	**32.0**	28.0
Fermat Point	24.0	24.0	30.0	30.0	28.0	**34.0**
Average	30.2	36.7	**44.7**	27.8	33.6	36.2
			Line Tasks			
Perpendicular	20.0	38.0	**46.0**	8.0	26.0	14.0
Parallel	22.0	28.0	30.0	20.0	**32.0**	**32.0**
Angle Bisector	28.0	36.0	**38.0**	28.0	28.0	24.0
Arc Connect	12.0	56.0	**68.0**	20.0	20.0	12.0
Perpendicular Bisector	22.0	20.0	40.0	16.0	30.0	**58.0**
Circle Tangent Line	22.0	20.0	**26.0**	22.0	20.0	22.0
Circle Tangent Point	18.0	16.0	**24.0**	22.0	18.0	22.0
Ray Reflection	28.0	30.0	**32.0**	**32.0**	18.0	26.0
Average	21.5	30.5	**38.0**	21.0	24.0	26.3
			Shape Tasks			
Square Outlier	54.0	44.0	54.0	52.0	54.0	**86.0**
Parallelogram	24.0	28.0	32.0	24.0	30.0	**36.0**
Right Triangle	30.0	14.0	16.0	38.0	20.0	**60.0**
Isosceles Trapezoid	36.0	**42.0**	36.0	24.0	26.0	20.0
Average	36.0	32.0	34.5	34.5	32.5	**50.5**
Overall Average	28.0	33.4	**40.2**	26.5	29.7	35.1

Here, "Binarize" uses a fixed threshold of 245. Pixels with intensity greater than this threshold are set to white (255), and others to black (0).

Evaluation For Sora-2, we manually evaluate the performance on each of the 10 tasks, based on the selected "best" frames (detailed in Sec. 2.1.2).

For the VLMs, we employ a rule-based evaluation by directly comparing their final answers with the ground truth answer for each test sample. For five of the 10 tasks, we provided multiple-choice options to reduce answer diversity and simplify evaluation. These five tasks are: Task 5 (Color Gradient Perception & Application) and the four shape-drawing tasks (Tasks 3, 4, 6, 10), all of which are illustrated in Fig. 13 above. For the other tasks, no multiple-choice options are provided. Detailed prompts are shown in Sec. G.1.2.

Results in Table We present the detailed evaluation results of visual puzzles in Tab. 10.

E.4. Text-Centric Tasks

Human Alignment Check for Evaluation To validate our evaluation method, we manually checked 173 cases sampled from the results of text-centric tasks. The rates at which the model correctly assessed the responses are 89.6% for video (last frame) and 97.7% for audio (transcribed answer), showing a relatively high level of consistency.

F. Supplementary Analysis and Results

This section provides additional analyses and experimental results that complement the main findings. We present details on data leakage analysis, output modality experiments, reasoning process categorization, and manual evaluation results.

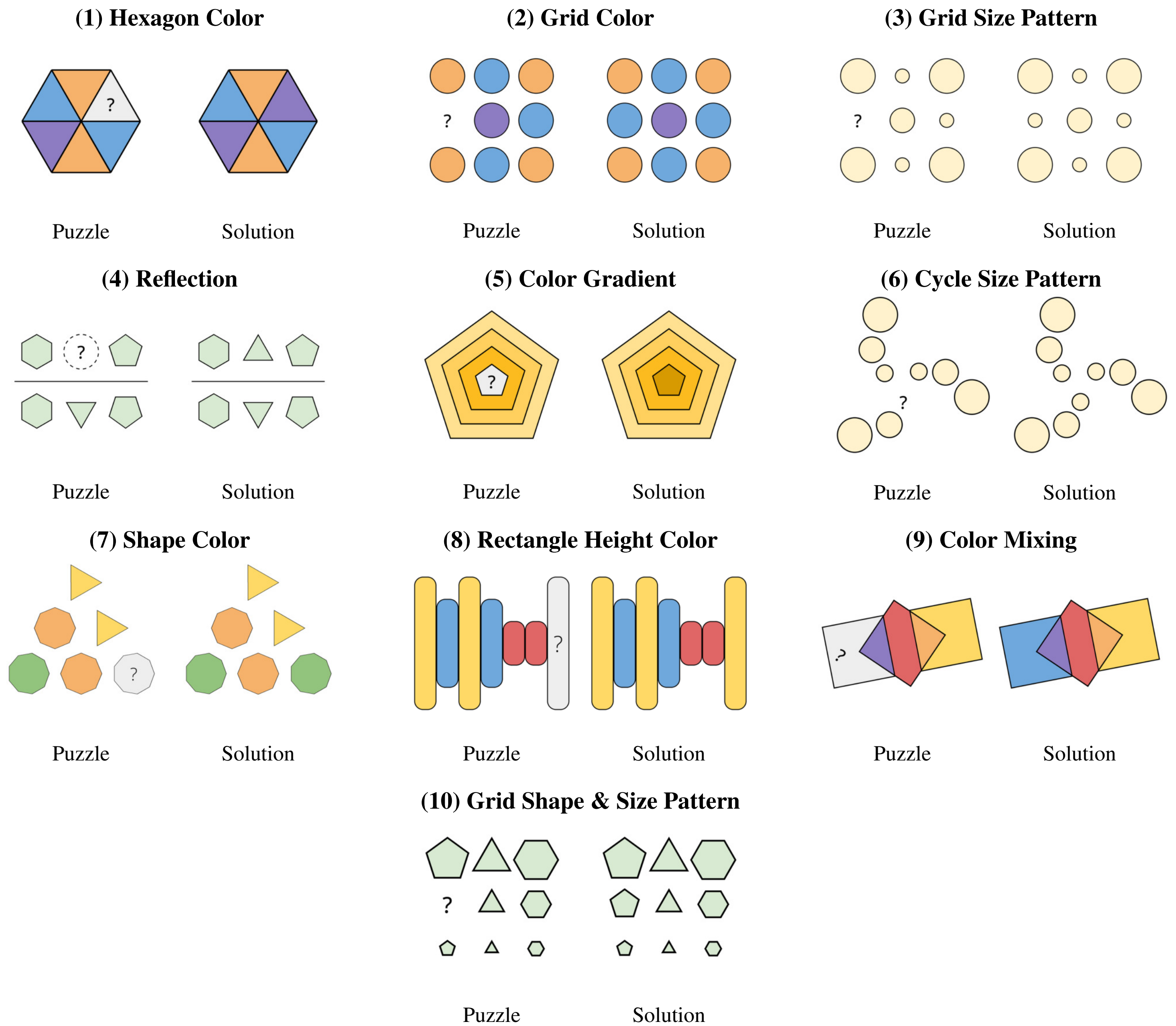


Figure 13. An overview of the 10 visual puzzle tasks evaluating inductive reasoning. The tasks are ordered by category: **Symmetry Tasks** (1)-(4), **Gradient Tasks** (5)-(6), and **Compositionality Tasks** (7)-(10). Each task displays a puzzle example and its solution.

F.1. Data Leakage Analysis

As mentioned in Sec. 3.2.1, we create new math evaluation problems to investigate potential data leakage as the reason for Sora-2's strong performance on text-centric tasks.

For each problem that we sampled from GSM8K [3] and MATH-500 [4], we used an LLM to derive a similar problem **with different numerical values** and possibly different contextual details while maintaining the overall difficulty. Qwen3-235B-A22B-Thinking [39] and Gemini 2.5 Pro [5] are used to derive the GSM8K and MATH-500 variants, respectively, with the prompts shown below.

Prompt for adapting the GSM8K problems

Given a grade school math problem and its solution, derive a new problem that is similar in the underlying problem-solving structure but with different numbers and, if possible, with different context and way of expression.
Ensure the new problem is solvable with an integer answer and maintains the same level of difficulty. Provide the solution to the derived problem using the same style and format as the original. Enclose your problem in <problem>and </problem>tags, and

Table 10. Accuracy (%) on the visual puzzle tasks. * represents that **multiple-choice options** are provided for the **VLMs** due to evaluation need, as detailed in Sec. E.3.1. Sora-2 is not provided with multiple-choice options across all 10 tasks.

Task	Sora 2	Gemini 2.5 Pro	GPT-5 high	Claude Sonnet 4.5
Symmetry Tasks				
Hexagon Color Pattern Match.	96.0	98.0	100.0	92.0
Grid Color Pattern Match.	94.0	94.0	100.0	100.0
Grid Size Pattern Match.*	85.4	87.5	95.8	62.5
Reflection Recognition & Application*	52.0	100.0	98.0	66.0
Average	81.9	94.9	98.5	80.1
Gradient Tasks				
Color Gradient Perception & Application*	45.8	83.3	35.4	93.8
Cycle Size Pattern Match.*	58.0	84.0	98.0	46.0
Average	51.9	83.7	66.7	69.9
Compositionality Tasks				
Color Mixing Perception & Application	56.0	56.0	100.0	86.0
Shape Color Pattern Match.	66.0	54.0	82.0	88.0
Rectangle Height Color Match.	44.0	58.0	60.0	54.0
Grid Shape & Size Pattern Match.*	64.0	100.0	98.0	100.0
Average	57.5	67.0	85.0	82.0
Overall Average	66.2	81.5	86.8	78.8

your solution in <solution>and </solution>tags.
Original Problem and Solution:
Problem: {original_problem}
Solution: {original_solution}

Prompt for adapting the MATH-500 problems

Given a math problem, derive a new problem that is similar in the underlying problem-solving structure but with different numbers and, if possible, with different context and way of expression.
Ensure the new problem is solvable and the complexity of its final answer is also similar to the original answer. For example, if the original answer is a simple integer (without any need to round), the new answer should also be a simple integer (also without any need to round).
Maintain the same level of difficulty.
Carefully analyze the original problem, think about the underlying structure, and carefully design the new problem to meet all the requirements above.
Provide the derived problem and a detailed solution to this new problem.
Enclose your problem in <problem>and </problem>tags, and your solution in <solution>and </solution>tags, and the final answer in <answer>and </answer>tags.

Original Problem and Answer:
Problem: {original_problem}
Answer: {original_answer}

F.2. Output Modality Analysis

To explore how output form affects Sora-2 performance, we designed the Arc Connect puzzle, which requires determining which right arc connects to the left arc to form part of a circle. An example of Sora-2 solving an Arc Connect puzzle is shown in Fig. 14. The evaluation methods of Sora-2 on Arc Connect puzzle are defined as follows:

- **Audio Option:** The prompt instructs model to speak out the option. Audio is extracted from generated video and transcribed to find the audio option.
- **Last Frame Option:** Last frame is extracted from the video. An evaluation program checks which right arc is connected to the left arc. If only one right arc is connected, the option is that option letter (“A” to “E”).
- **Major Frame Option:** For every 5 frames in the video, one frame is extracted and fed to the evaluation program, getting option of this frame. Major Frame Option is the majority vote result of all chosen frames.

F.3. Prompt Rewriting in Wan2.5: Case Study

As discussed in Sec. 3.2.3, Wan2.5’s text-centric reasoning ability is almost entirely attributed to its prompt rewriter model. Here we provide a concrete example demonstrat-

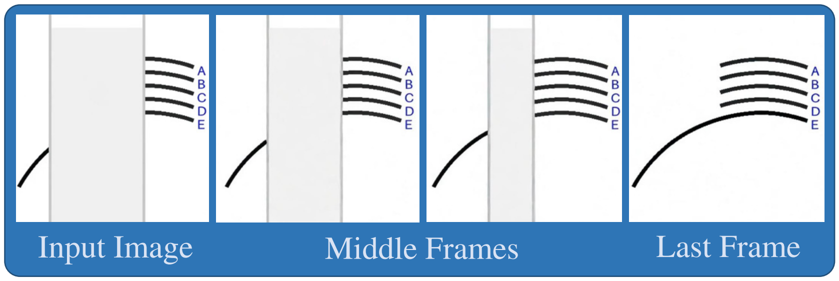


Figure 14. **Sora-2 solving an Arc Connect puzzle.** Prompt: *"One arc on the left continues across the masked band to one of the arcs on the right. Which labeled arc matches? Remove the masked band quickly while keeping the arcs still. Speak out the answer in phonetic alphabet. In portrait. Static Camera. No zoom."* Sora-2 successfully removes the band. Details: Section F.2.

ing how prompt rewriting transforms a reasoning task into explicit step-by-step visual instructions for the video generation component.

F.3.1. Example: GSM8K Problem

Original Prompt (Without Explicit Rewriting)

Solve the problem step by step on the given whiteboard. Give the final answer by writing "The answer is ... (final answer)". No oral explanation was provided during the written process of solving the problem, but the final answer was stated orally in the end.

Problem: There are 6 girls in the park. If there are twice the number of boys in the park, how many kids are in the park?

Rewritten Prompt (After Prompt Rewriting)

The problem is presented on a whiteboard. A hand writes 'Girls = 6'. Then, 'Boys = 2 × 6 = 12' appears. Next, 'Total kids = 6 + 12 = 18' is written. Finally, 'The answer is 18' is written on the board. A voice states: 'The answer is 18'.

The rewritten prompt explicitly specifies the solution steps and visual elements to be generated, effectively solving the problem before video generation. This transformation explains why disabling the prompt rewriter leads to nearly zero accuracy (Tab. 6).

F.3.2. Visual Comparison

Fig. 15 shows the visual outputs generated by Wan2.5 with and without prompt rewriting enabled.

F.4. Reasoning Process Categorization

In Sec. 3.2.2, we analyzed Sora-2's reasoning processes for text-centric tasks. Here we provide the detailed categorization scheme and examples.

Table 11. Accuracy (%) on the ARC-AGI-2 task. We display each sample in an image and send it to Sora-2 and VLMs. Details: Sec. 2.1.3

Task	Sora-2	Gemini 2.5 Pro	GPT-5 high	Claude Sonnet 4.5	Grok-4
ARC-AGI-2	1.3	1.9	0.5	5.3	2.7

F.4.1. Category Definitions

We categorize the solution process into five categories:

1. **Completely Correct:** The solution has a clear and correct process without any errors.
2. **Logic Correct with Writing Errors:** The solution contains expressional mistakes, but the overall logic is identifiable and correct.
3. **Unreadable or Incorrect Logic:** The writing is too disorganized or contains too many errors to discern the reasoning, or it exhibits clear logical mistakes or major omissions.
4. **Missing Solution Process:** Necessary steps are absent; apart from the final answer, the response is blank or contains only meaningless scribbles (i.e., lines, circles, etc).
5. **Process Unnecessary:** The problem itself does not require a written process to solve.

F.5. Evaluation Results of ARC-AGI-2

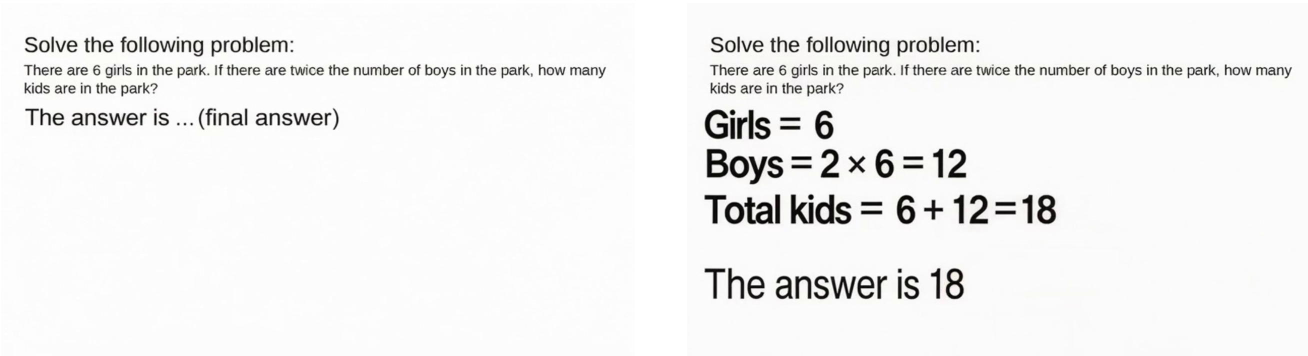
We show the results in Tab. 11.

F.6. Manual Evaluation of ARC-AGI-2

To provide a more fine-grained assessment of Sora-2's performance on ARC-AGI-2 beyond binary correctness, we manually evaluated 100 randomly selected samples and categorized them into different quality levels.

G. Prompts

In this section, we list all prompts used for evaluation.



(a) Without prompt rewriting (`prompt_extend=false`): The model generates meaningless or incorrect content.

(b) With prompt rewriting (`prompt_extend=true`): The model correctly displays solution steps specified in the rewritten prompt (Sec. F.3.1).

Figure 15. Visual comparison of Wan2.5's outputs with and without prompt rewriting on the same GSM8K problem. The dramatic difference demonstrates that the reasoning capability resides in the prompt rewriter rather than the video generation model itself.

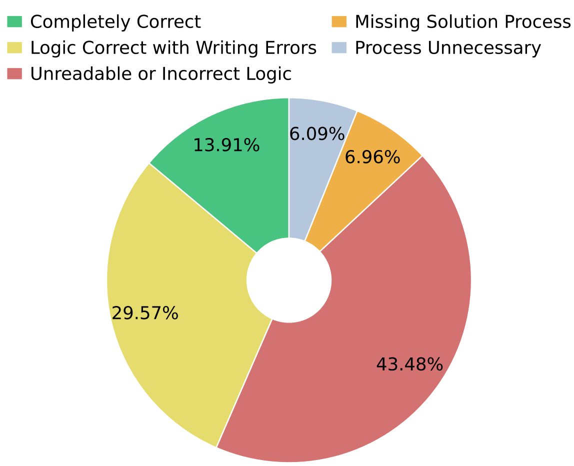


Figure 16. Distribution of reasoning process categories for correct answers. The prevalence of "Unreadable or Incorrect Logic" (43.48%) highlights Sora-2's difficulty in generating coherent reasoning.

Category	Fully Correct	Mostly Correct	Partially Correct	Wrong	
				Did Nothing	Others
Count	3.0	14.0	28.0	42.0	13.0

Table 12. Manual evaluation of 100 randomly chosen ARC-AGI-2 samples.

G.1. Vision-centric Tasks

G.1.1. Eyeballing Puzzles

We list prompts used for each puzzle as below.

Prompt for Arc Connect

Sora-2 prompt: *One arc on the left continues across the masked band to one of the arcs on the right. Which labeled arc matches? Remove the masked band quickly while keeping the arcs still. Speak out the answer in phonetic alphabet. In portrait. Static Camera. No zoom.*

VLMs prompt: *One arc on the left continues across the masked band to one of the arcs on the right. Which labeled arc matches? Answer an option in A-E.*

Prompt for Angle Bisector

Sora-2 prompt: *Draw a black line bisecting the angle. Speak out which option is on the bisector using phonetic alphabet and mark that red. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is on the bisector of the angle? Answer an option in A-E.*

Prompt for Circle Center

Sora-2 prompt: *Mark the center of the circle red. Speak out which option is the center using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the center of the circle? Answer an option in A-E.*

Prompt for Circle Tangent Line

Sora-2 prompt: *Draw a black line tangent to the circle at the highlighted point. Speak out which option lies on this tangent line in phonetic alphabet and mark that red. In portrait, no zoom, no pan.*

VLMs prompt: *Which option lies on the line that is tangent to the circle at the highlighted point? Answer an option in A-E.*

Prompt for Circle Tangent Point

Sora-2 prompt: *Draw the tangent line from the external point to the circle in black. Paint the point of tangency red. Speak out which option is the point using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the tangent point on the circle from the external point? Answer an option in A-E.*

Prompt for Circumcenter

Sora-2 prompt: *Mark the circumcenter of the triangle red. Speak out which option is the circumcenter using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the circumcenter of the triangle? Answer an option in A-E.*

Prompt for Fermat Point

Sora-2 prompt: *Find the Fermat point of the triangle. Mark the point red. Speak out which option is the Fermat point using the phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the Fermat point of the triangle? Answer an option in A-E.*

Prompt for Incenter

Sora-2 prompt: *Mark the incenter of the triangle red. Speak out which option is the incenter using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the incenter of the triangle? Answer an option in A-E.*

Prompt for Isosceles Trapezoid

Sora-2 prompt: *Find the fourth vertex that completes the isosceles trapezoid. Mark the fourth vertex red. Speak out which option is the fourth vertex using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the fourth vertex of the isosceles trapezoid? Answer an option in A-E.*

Prompt for Midpoint

Sora-2 prompt: *Connect the two large circles and mark the midpoint as red. Speak out which option is the midpoint using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the midpoint of the two circles? Answer an option in A-E.*

Prompt for Orthocenter

Sora-2 prompt: *Find the orthocenter (intersection of altitudes) of the triangle and mark it red. Speak out which option is the orthocenter using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the orthocenter of the triangle? Answer an option in A-E.*

Prompt for Parallel

Sora-2 prompt: *Draw a black line through the small circle and parallel to the existing line. Speak out which option is on the new line using phonetic alphabet and mark that red. In portrait, no zoom, no pan.*

VLMs prompt: *Draw a line through the small circle and parallel to the existing line, which option is on it? Answer an option in A-E.*

Prompt for Parallelogram

Sora-2 prompt: *Draw a black parallelogram with two sides given. Mark the fourth vertex red. Speak out which option is the fourth vertex using phonetics alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the fourth vertex of the parallelogram with two sides given? Answer an option in A-E.*

Prompt for Perpendicular

Sora-2 prompt: *Draw a black line perpendicular to the existing line and passing the small circle. Speak out which option is on the line using phonetic alphabet and mark that red. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is on the line perpendicular to the black line and passing the small circle? Answer an option in A-E.*

Prompt for Perpendicular Bisector

Sora-2 prompt: *Draw a black line that is the perpendicular bisector of the segment between the two small circles. Speak out which option is on the line using phonetic alphabet and mark that red. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is on the perpendicular bisector of the segment connecting the two small circles? Answer an option in A-E.*

Prompt for Ray Intersection

Sora-2 prompt: *Extend the three black lines and mark the intersection point as red. Speak out which option is the intersection point using phonetics alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the intersection point of the three lines? Answer an option in A-E.*

Prompt for Ray Reflection

Sora-2 prompt: *Draw the ray of light starting from the small circle and reflecting off the line in black. Speak out which option the reflected ray will pass through using phonetic alphabet and mark it red. In portrait, no zoom, no pan.*

VLMs prompt: *A ray of light starts from the small circle and reflects off the line. Which option will the reflected ray pass through? Answer an option in A-E.*

Prompt for Point Reflection

Sora-2 prompt: *Reflect the small circle across the line. Mark the reflection red and speak out which option is the reflected point using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the reflection of the small circle across the line? Answer an option in A-E.*

Prompt for Right Triangle

Sora-2 prompt: *Out of the 5 points, 3 form a right-angled triangle. Mark the vertex with the right angle in red. Speak out which option is the right-angle vertex using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the vertex of the right angle, given that exactly three of the five options form a right-angled triangle? Answer an option in A-E.*

Prompt for Square Outlier

Sora-2 prompt: *Four of the five options form a square. Mark the fifth point red. Speak out which option is the fifth point using phonetics alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Four of the five options form a square. Which option is the fifth point? Answer an option in A-E.*

Prompt for Triangle Center

Sora-2 prompt: *Mark the center of the triangle red. Speak out which option is the center using phonetic alphabet. In portrait, no zoom, no pan.*

VLMs prompt: *Which option is the center of the triangle? Answer an option in A-E.*

G.1.2. Visual Puzzles

Prompt for Tasks 1, 2, 3, 4 and 6

Sora-2 prompt: *What is the missing color of the part denoted with a question mark? This part should be completely filled with the correct color while the other parts should be unchanged. The question mark disappears. Then nothing happens and the scene remains static. Do not zoom in or out, or change the positions of the shapes.*

VLMs prompt: *What is the missing color of the part denoted with a question mark?*

Prompt for Task 5 (Color Gradient Perception & Application)

Sora-2 prompt: *What is the missing color of the part denoted with a question mark? This part should be completely filled with the correct color (not white or the original grey) to match the pattern in the image while the other parts should be unchanged. The question mark disappears. Then nothing happens and the scene remains static. Do not zoom in or out, or change the positions of the shapes.*

VLMs prompt: *What is the missing color of the part denoted with a question mark? Options: ...* (Four options.)

Prompt for Task 7 (Grid Size Pattern Matching) and 8 (Cycle Size Pattern Matching)

Sora-2 prompt: *What is the size of the missing part denoted with a question mark? This part should be replaced with the correct circle while the other circles should be unchanged. The question mark disappears. Then nothing happens and the scene remains static. Do not zoom in or out, or change the positions of the shapes.*

VLMs prompt: *What is the size of the missing circle denoted with a question mark? Options: small, medium, large* (The three options are randomly shuffled.)

Prompt for Task 9 (Grid Shape & Size Pattern Matching)

Sora-2 prompt: *What is the size of the missing part denoted by a question mark? This part should be replaced with the correct shape while the other shapes should be unchanged. The question mark disappears. Then nothing happens and the scene remains static. Do not zoom in or out, or change the positions of the shapes.*

VLMs prompt: *What is the size of the missing part denoted by a question mark? Options: small, medium, large.* (The three options are randomly shuffled.)

Prompt for Task 10 (Reflection Recognition & Application)

Sora-2 prompt: *What is the missing shape denoted by a question mark? The question mark area should be replaced with the correct shape while the other shapes should be unchanged. The question mark disappears. Then nothing happens and the scene remains static. Do not zoom in or out, or change the positions of the shapes.*

VLMs prompt: *What is the missing shape denoted by a question mark? Options: triangle, square, pentagon, hexagon.* (The four options are randomly shuffled.)

G.2. Text-centric Tasks

G.2.1. Generation

Prompt for problems from GSM8K, MATH-500, AIME and GPQA-diamond

Solve the problem step by step on the given whiteboard. No oral explanation was provided during the written process of solving the problem, but the final answer was stated orally in the end, which is also clearly written.
Problem: {problem}

Prompt for problems from BBH, MMLU, MMLU-Pro and SuperGPQA-easy

A short video explaining a multiple-choice question.

Visual Setup:
*- **Background:** A solid, pure white background throughout the entire video.*
*- **Layout:** Split-screen layout.*
- Clearly displays the question and multiple-choice options. Use a large, clean, and easy-to-read font.
- No presenter or other irrelevant content to the question.
- The question is displayed at the top center with "Question:" as the title
- Multiple-choice options (A, B, C, etc.) are listed below - A "Correct Answer: _____" line appears at the bottom with appropriate spacing from the edge
- All text uses clear, easy-to-read fonts
Content to Display:
Question:
{question}
Correct Answer: _____ (fill this in after explanation)
Requirements:

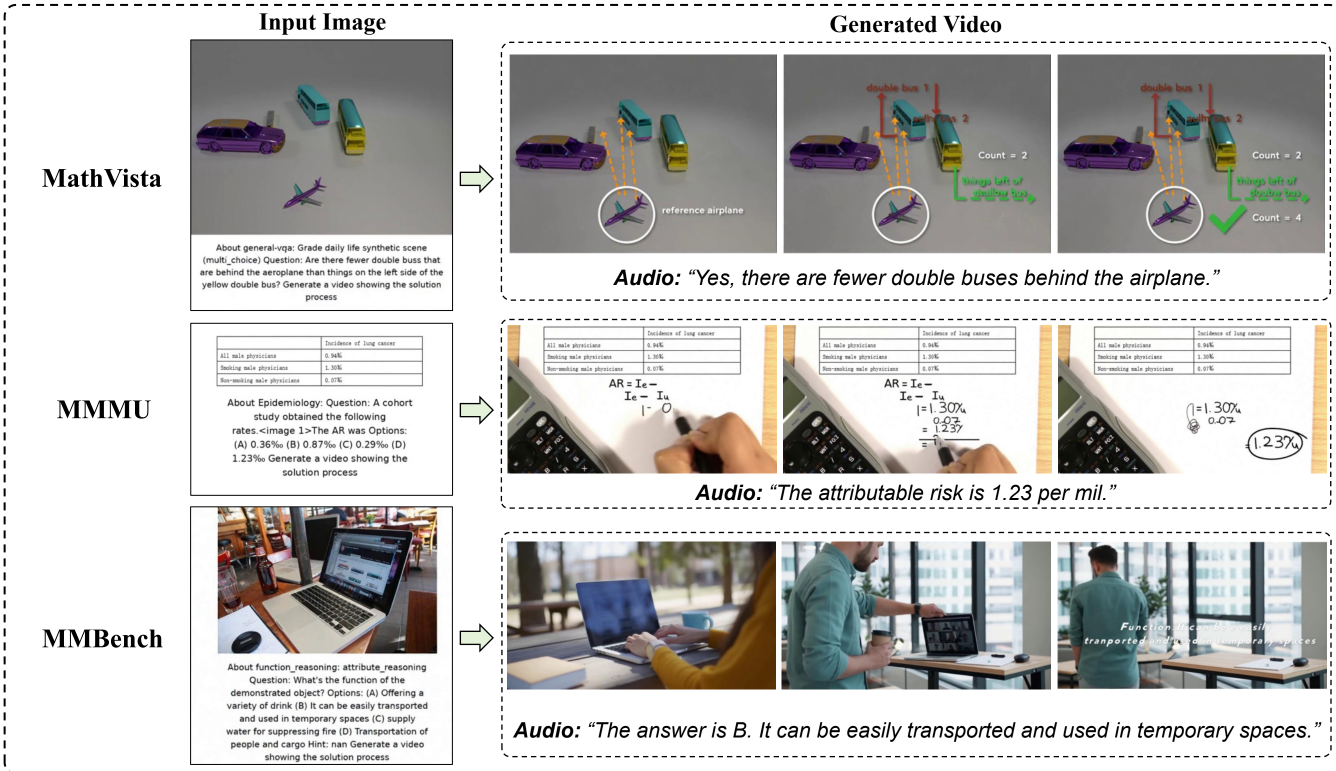


Figure 17. Sora-2 solves the multimodal reasoning questions of the text-centric tasks. The input image contains the image of the original multimodal reasoning problem and the question. Similar to Fig. 6, a text prompt containing the question text is also input to the model.

- Directly state the correct answer through audio narration (e.g., "The correct answer is A" or "The answer is True")
- Fill in the correct answer in the "Correct Answer: _____" line on screen
- No need for explanation or reasoning - just clearly announce the answer
***Style & Tone:** Clear and articulate voice, professional tone, direct and concise.*

Prompt for problems from MathVista, MathVision, MMBench and MMMU

Question: {question}
Generate a video showing the solution process

G.2.2. Evaluation

For the text-centric tasks, we use GPT-4o [20] as the judge model to evaluate the answer from the video and the audio independently, with the prompts shown below. For audio transcription, we use OpenAI's whisper model (whisper-1) via its API.

Prompt for Evaluating the Answer from the Video

System prompt:

You are an expert answer checker for educational videos. Your task is to determine if an image (the last frame of a solution video) displays the correct answer to a given question.

Rules:
0. First, determine the visible answer from the image using this priority:
- If there is an explicit statement indicating the answer (e.g., "The answer is ..."), use that answer.
- Else, check for an answer marked by a symbol such as box, circle, underline, arrow, etc. If multiple positions are marked but show different results, respond 'no' immediately.
- Else, use the bottom-rightmost result in the image as the visible answer.
1. Compare the visible answer in the image with the provided correct answer
2. Be strict but reasonable - minor formatting differences are acceptable if the core answer is correct
3. For multiple choice questions, check if the correct

option (A, B, C, etc.) is clearly marked or highlighted
4. For numerical answers, check if the number matches (ignore minor formatting like "4" vs "4.0")
5. For text answers, check if the key content matches (ignore case sensitivity and minor punctuation)
6. You must respond with ONLY 'yes' or 'no', nothing else

User instruction prompt:

Question: {question}

Correct answer: {correct_answer}

Does the image show the correct answer?

(The last frame of the generated video is also provided for the model.)

Prompt for Evaluating the Answer from the Audio

System prompt:

You are an expert answer checker for educational video transcripts. Your task is to determine if an audio transcript from a solution video contains the correct answer to a given question.

Rules:
1. Check if the transcript explicitly states or clearly implies the correct answer
2. Be lenient with phrasing - the transcript may explain the answer in different words
3. For multiple choice questions, check if the correct option (A, B, C, etc.) is mentioned
4. For numerical answers, check if the number is stated (ignore surrounding explanation)
5. For text answers, check if the key concept is explained correctly
6. Common phrases like "the correct answer is...", "the answer is...", "it should be..." indicate the answer
7. You must respond with ONLY 'yes' or 'no', nothing else

User instruction prompt:

Question: {question}

Correct answer: {correct_answer}

Audio transcript: {transcript}

Does the transcript provide the correct answer?

H. Multimodal Reasoning Cases

We supplement the cases of Sora-2's solving multimodal reasoning questions of the text-centric tasks, as illustrated in Fig. 17.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[6] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

[7] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[8] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

[9] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

[10] Google. Veo 3. `https://aistudio.google.com/models/veo-3`, 2025. Accessed on March 28, 2026.

[11] Google DeepMind. Gemini 3 pro image - model card. Technical report, Google DeepMind, 2025. Published: November 20, 2025.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning ca-

pability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[13] Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. *arXiv preprint arXiv:2510.26802*, 2025.

[14] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[15] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models, 2025.

[16] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing advanced visual reasoning ability of large language models, 2024.

[17] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[18] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[19] MiniMax. Minimax hailuo 2.3: A new level of complex video performance & media agent. https://www.minimax.io/news/minimax-hailuo-23, 2025. Accessed: 2026-03-27.

[20] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[21] OpenAI. New ChatGPT images is here. https://openai.com/zh-Hans-CN/index/new-chatgpt-images-is-here/, 2025. Published: December 16, 2025.

[22] OpenAI. Learning to reason with llms, 2025. Accessed: 2025.

[23] OpenAI. OpenAI o3 and o4-mini System Card. Technical report, OpenAI, 2025. Accessed: 2025-11-01.

[24] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[25] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[26] Runway Research. Introducing Gen-3 Alpha: A New Frontier for Video Generation. https://runwayml.com/research/introducing-gen-3-alpha, 2024. Accessed on March 28, 2026.

[27] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

[28] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022.

[29] OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

[30] Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Game-rl: Synthesizing multimodal verifiable game data to boost vlms' general reasoning. *arXiv preprint arXiv:2505.13886*, 2025.

[31] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[32] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[33] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[34] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[36] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, et al. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

[37] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[38] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

[39] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen

Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[40] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms, 2025.

[41] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[42] Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis, 2025.

[43] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning, 2025.

[44] Yongheng Zhang, Xu Liu, Ruihan Tao, Qiguang Chen, Hao Fei, Wanxiang Che, and Libo Qin. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models, 2025.

[45] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024.

[46] Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, and Xuanjing Huang. Exploring the compositional deficiency of large language models in mathematical reasoning. *arXiv preprint arXiv:2405.06680*, 2024.