Thinking with Video

Video Generation as a Promising Multimodal Reasoning Paradigm

视频生成作为多模态推理新范式

CVPR 2026


Fudan University, OpenMOSS, SII

https://thinking-with-video.github.io

先前的推理范式

Thinking with Text

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

早期：通过提示词激发模型逐步推理

Table 2 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$

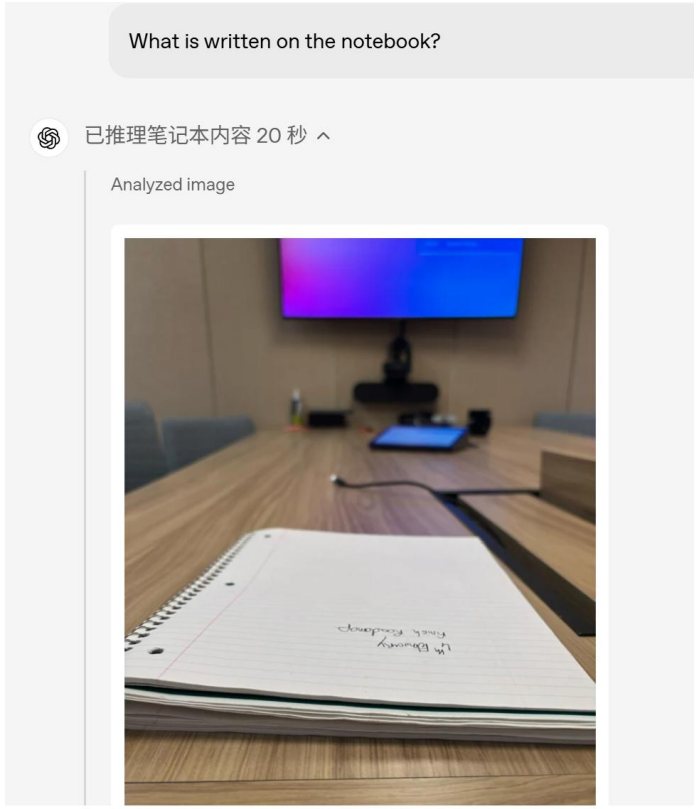
通过强化学习训练，激发模型更加强大、类人的长链推理能力（DeepSeek-R1）

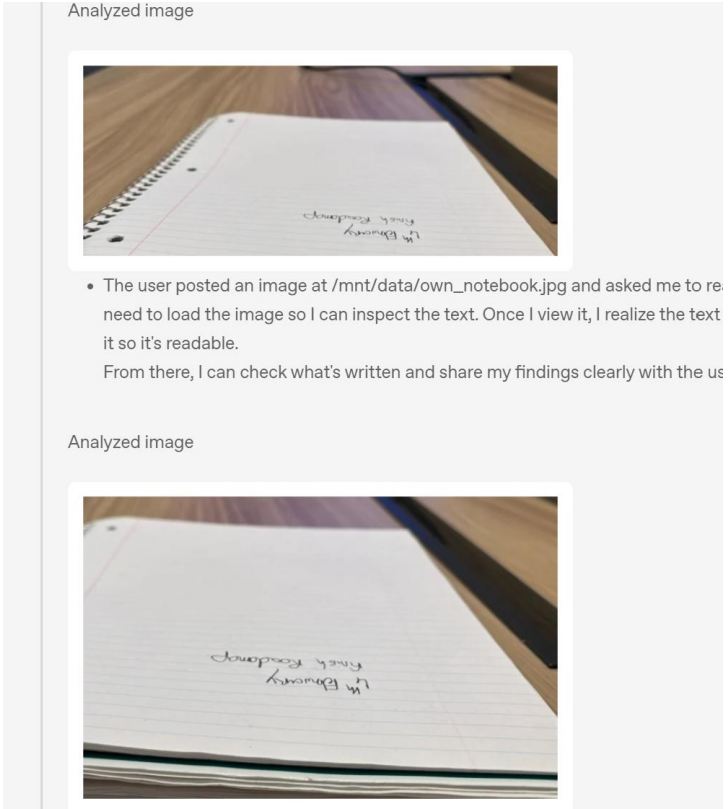
https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
https://arxiv.org/pdf/2501.12948

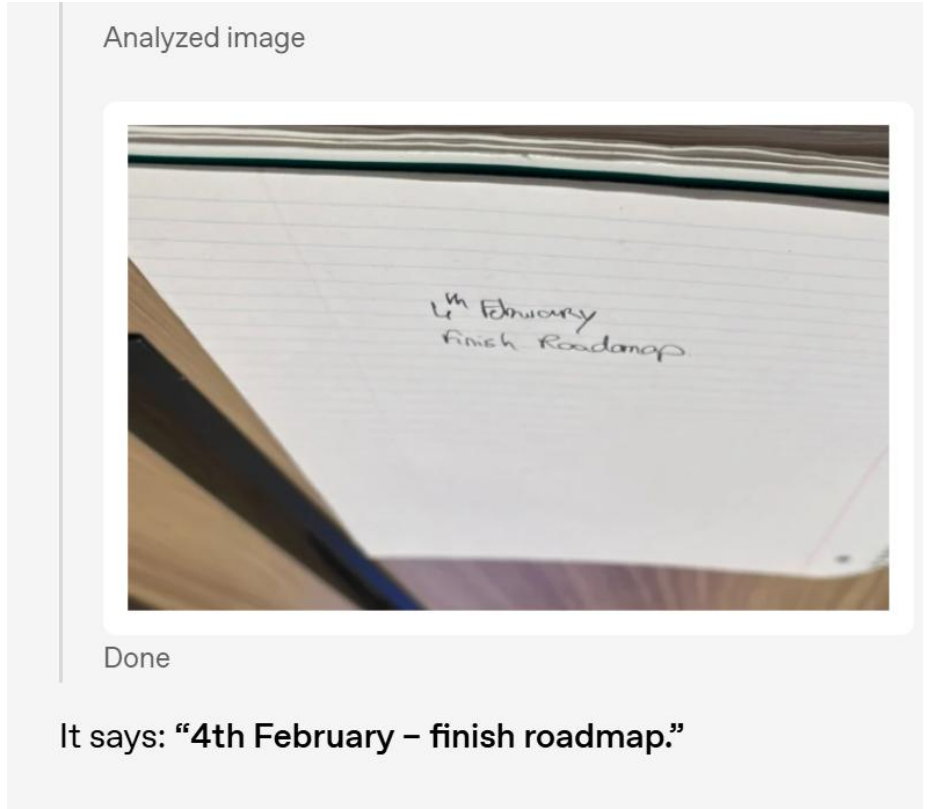
先前的推理范式

Thinking with Images

在 CoT 中加上生成的图像（如通过工具调用编辑图像），辅助 VLM 进行推理







OpenAI o3 和 o4-mini 能自主编辑图片辅助其进行推理

https://openai.com/zh-Hans-CN/index/thinking-with-images/

为什么需要新的推理范式

先前的推理范式

Thinking with Text

基于长文本 CoT 的推理

提升 LLM/VLM 的能力

Thinking with Images

在 CoT 中加上生成的图像

辅助 VLM 进行推理

问题：

- **对于多模态推理，Thinking with Text 和 Thinking with Images 是否有天然缺陷?**
- **人类的推理行为还有什么特点?**

为什么需要新的推理范式

Thinking with Text 和 Thinking with Images 的天然缺陷

1. 静态约束

- 图像只能捕捉单一时刻的信息
- 难以表达动态过程、时间变化与连续变换

2. 模态分离

- 文本与视觉仍被分开处理
- 缺少一种自然统一二者的推理载体

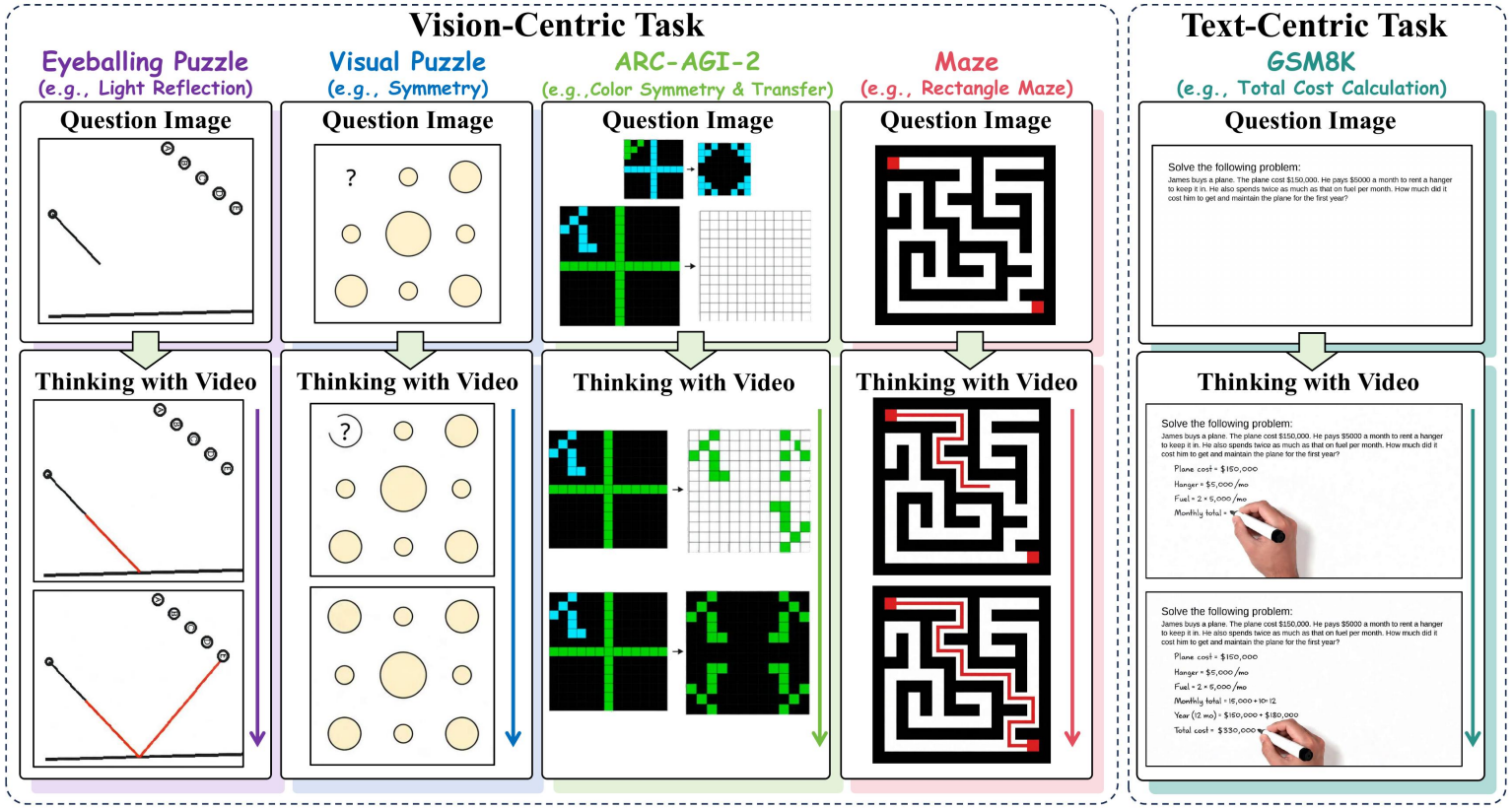
需要一种更自然的时序媒介

需要能统一文本与视觉的载体

Thinking with Video

将视频生成作为一种多模态推理范式

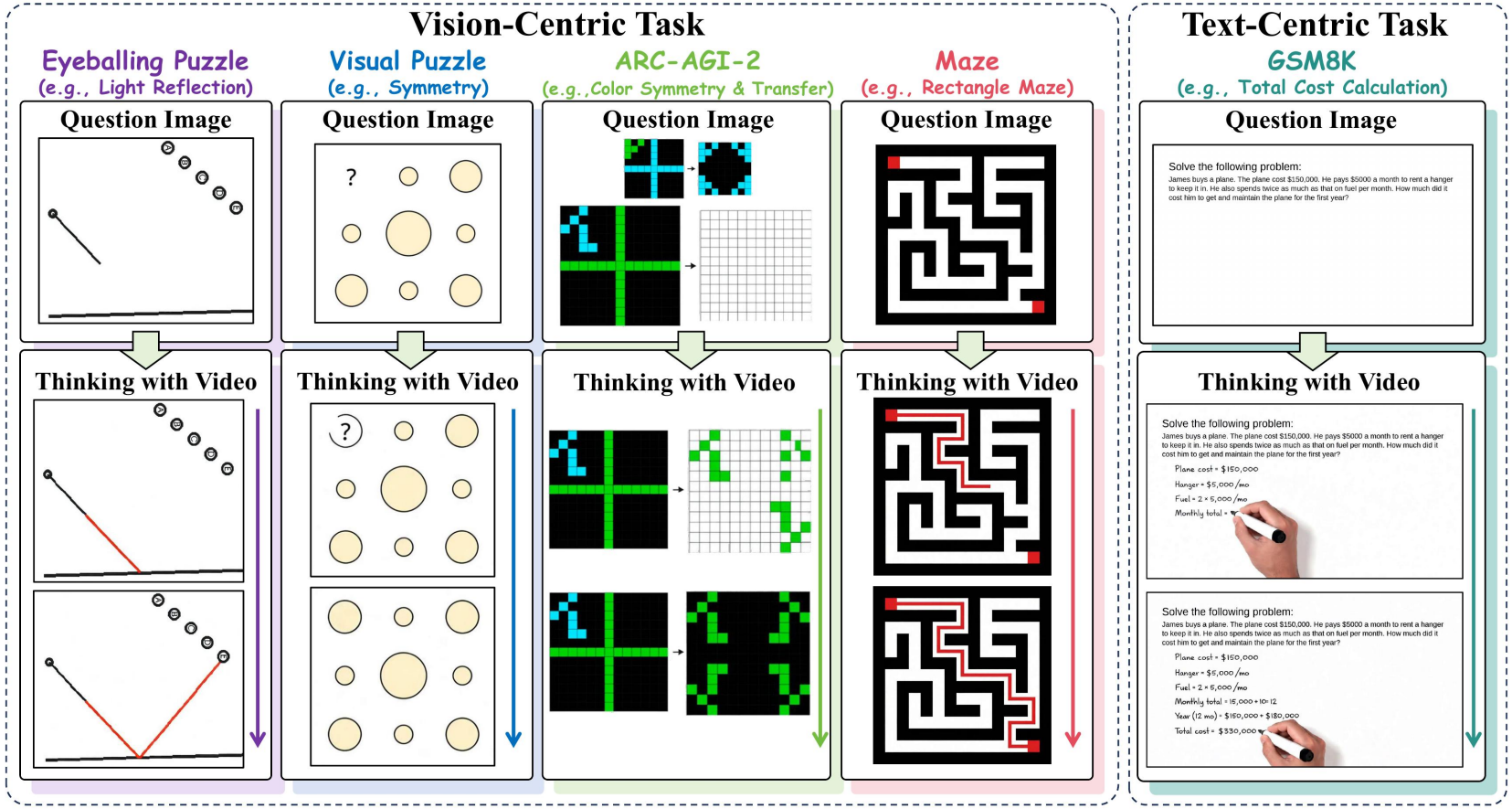
- ✓ 视频生成模型可进行**绘制、想象、模拟**，有助于解决相关视觉推理问题
- ✓ 视频帧可同时**承载文本**，进而也有望完成一些文本推理问题



以视频帧为统一媒介进行多模态推理

测试基准：VideoThinkBench

覆盖以视觉中心的任务 & 以文本为中心的任务



Vision-Centric Task

基于绘制与想象

Text-Centric Task

基于文本书写

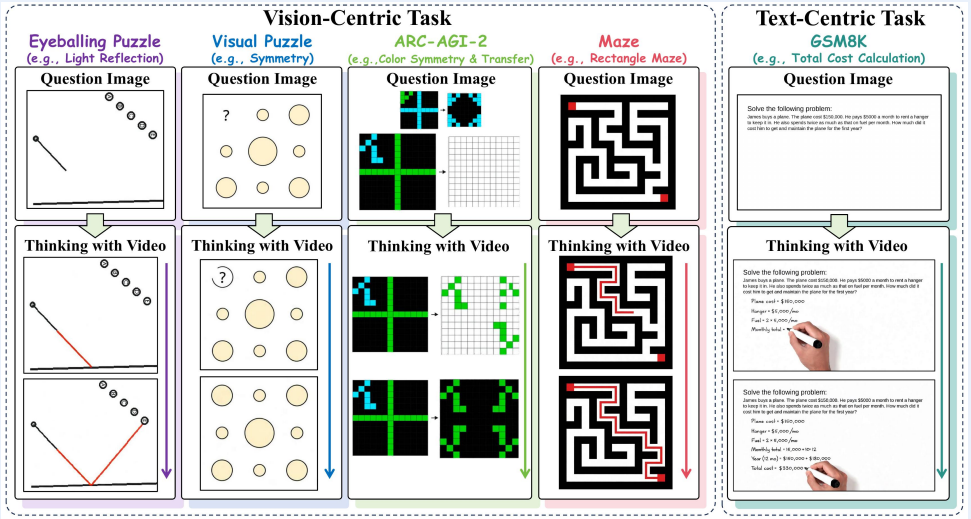
Goal

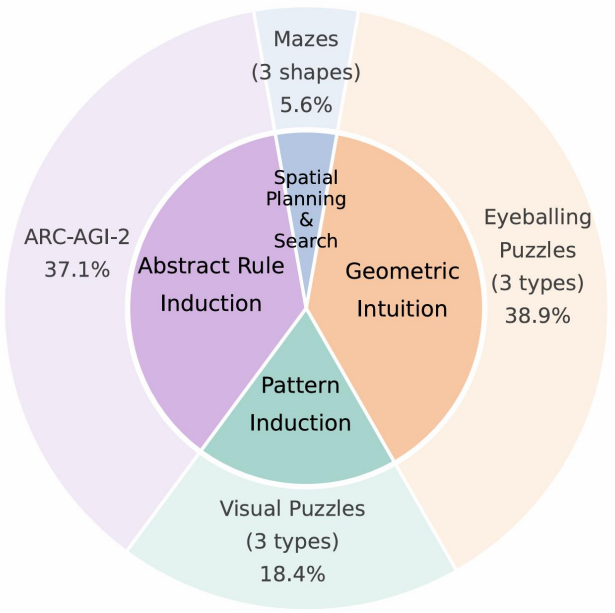
综合评估视频推理能力

VideoThinkBench 的能力层级

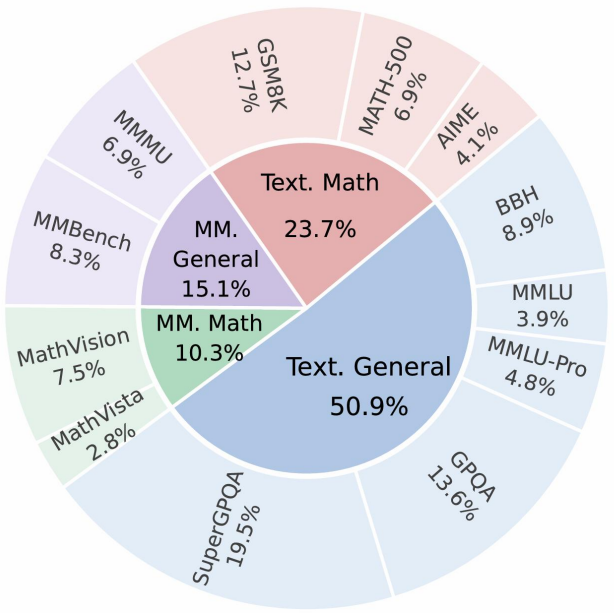
五类核心推理能力

1. 几何直觉：Eyeballing Puzzles
2. 视觉归纳：Visual Puzzles
3. 抽象规则归纳：ARC-AGI-2
4. 空间规划与搜索：Mazes
5. 语言概念理解与推理：Text-Centric Tasks





(a) Vision-centric reasoning tasks



(b) Text-centric reasoning tasks

从视觉结构推理，延伸到文本的语言推理

总体评测设置

Sora-2 与 SOTA VLM 对比

视觉任务

- 数据由程序生成
- 多数任务能自动验证，可精准评测

文本任务

- 改编自已有的文本测试基准
- 评测视频帧上以及语音中的视频模型的回答

对比模型

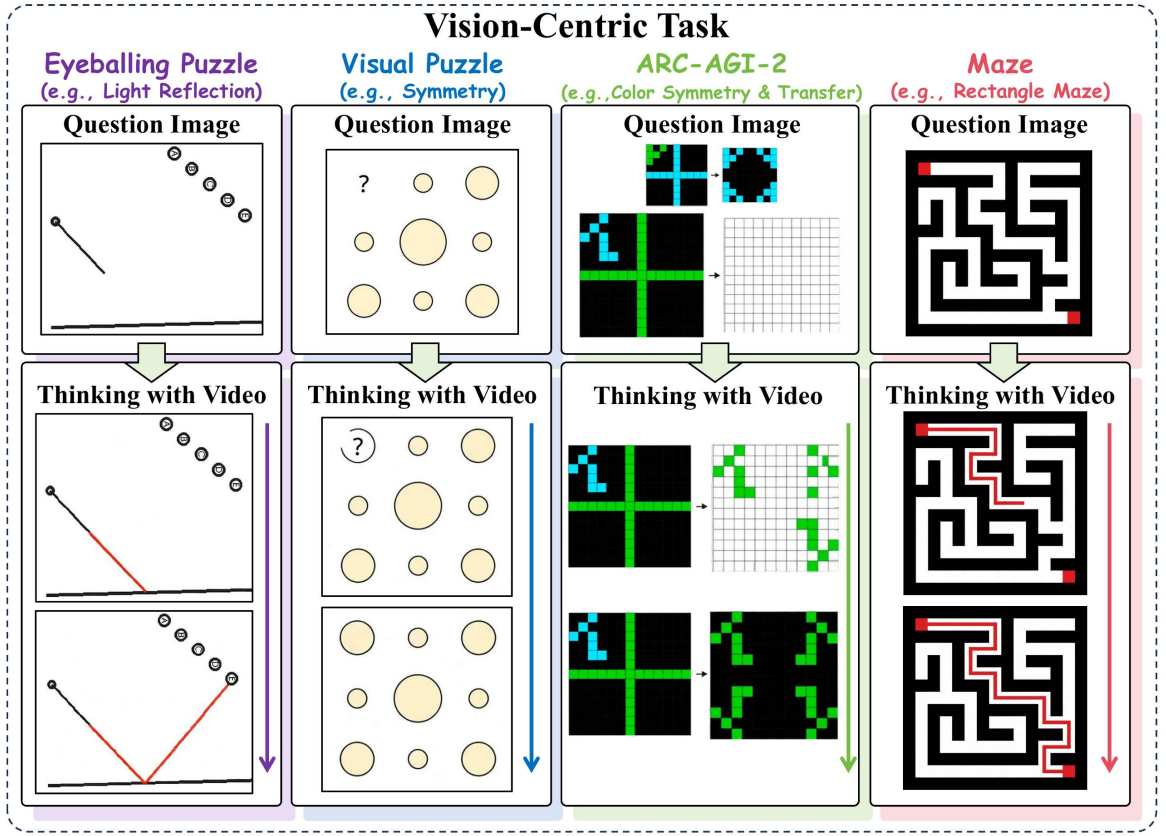
- Gemini 2.5 Pro
- GPT-5 high
- Claude Sonnet 4.5

核心问题：视频生成模型是否具有推理能力？相比 SOTA VLM 是否有优势？

Vision-Centric Reasoning

视频模型是否真的会“看着推理”

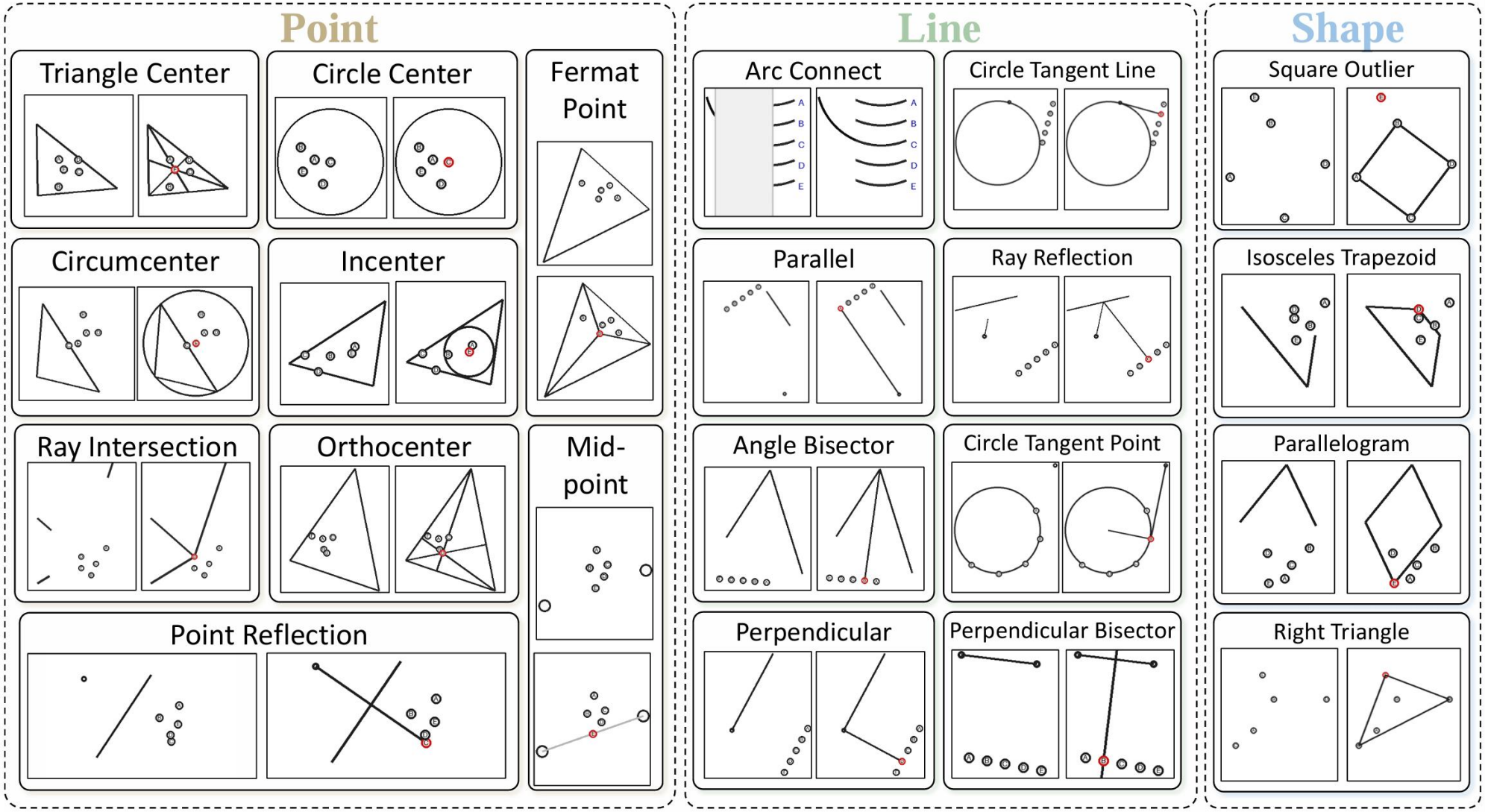
- 视觉任务要求模型在画面中构造、追踪和修改结构
- 这类能力更接近人类的想象与草稿推理



绘制 + 想象，是 Thinking with Video 的关键优势

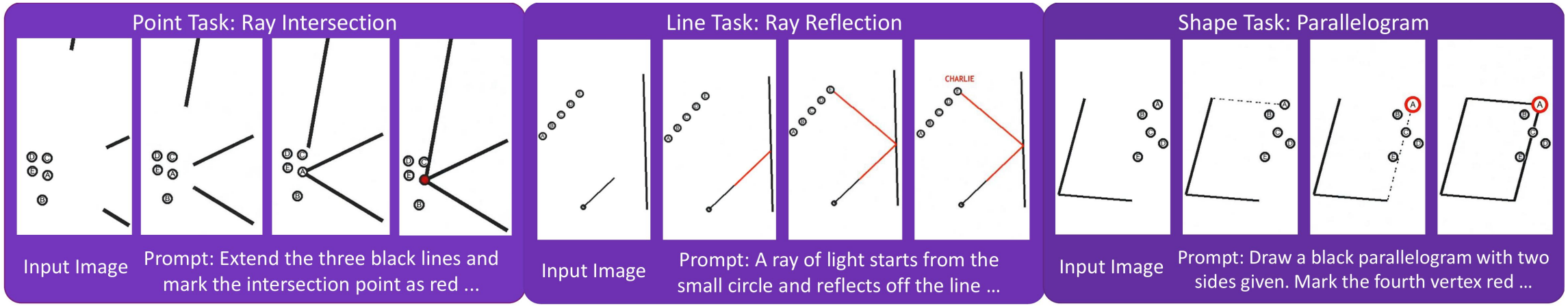
Eyeballing Puzzles：几何直觉

通过绘制完成几何推理



Eyeballing Puzzles：几何直觉

通过绘制完成几何推理



任务

21 类可验证几何选择题
共 1050 个样本

视频推理

画线（延长图线等等）
并标记答案

评估

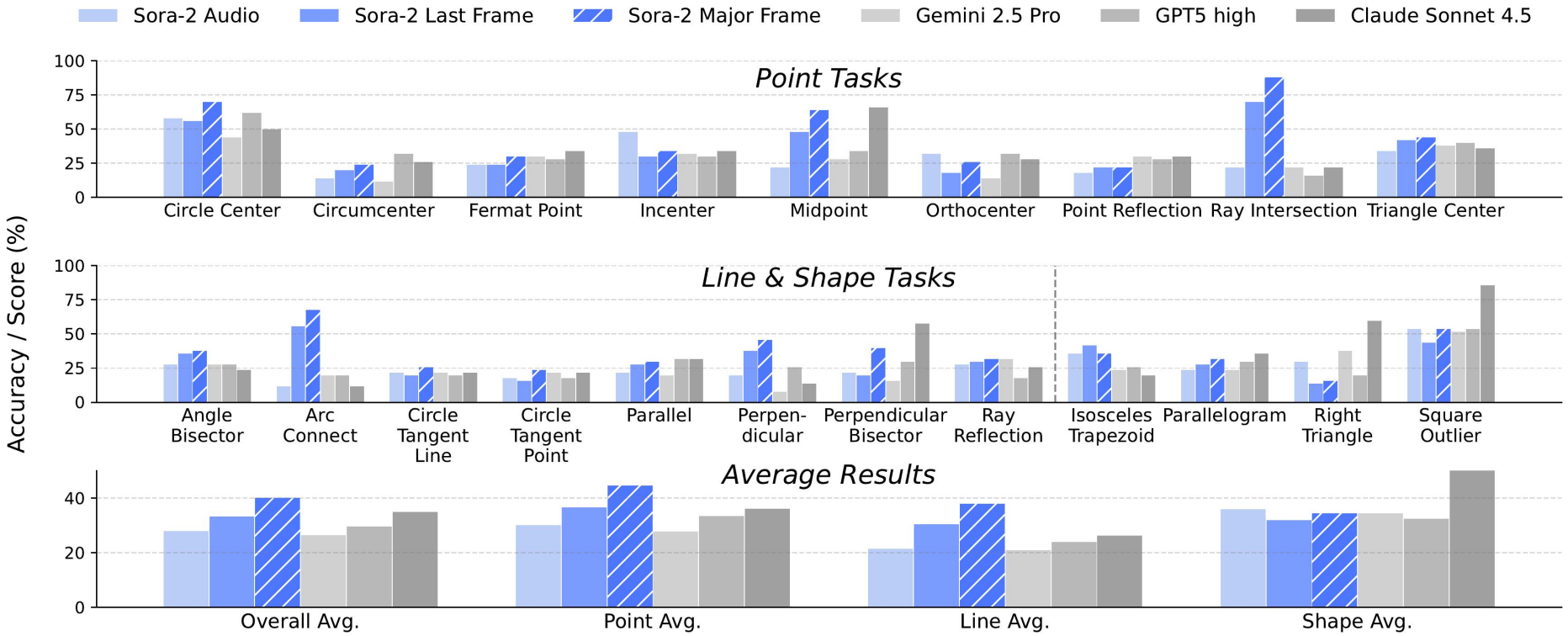
Audio / Last Frame
/ Major Frame

Major Frame：

- 不只是看最后一帧的结果，而是利用视频过程，通过多帧投票确定答案
- 好处：利用整个视频过程，避免最后一帧的噪声

Eyeballing Puzzles：结果

Sora-2 展现强大几何推理能力，总体超越了三个 SOTA VLM

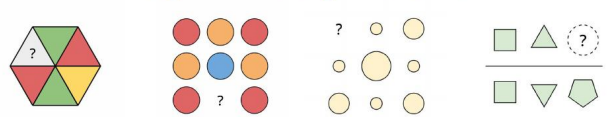


Sora-2 Major Frame：40.2%，高于 Claude 35.1%、GPT-5 29.7%、Gemini 26.5%

Visual Puzzles：视觉归纳推理

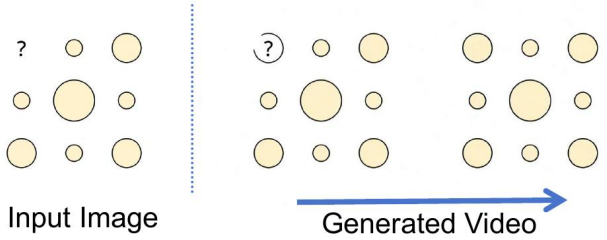
关于颜色、形状、尺寸进行归纳推理（inductive reasoning）

Symmetry Tasks



Example: *Grid Size Pattern Matching*
(size symmetry in a grid)

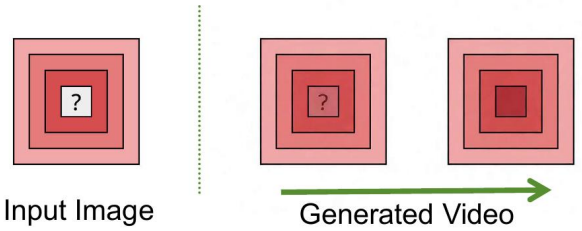
Prompt Example: *The question mark area should be replaced with the correct shape.*



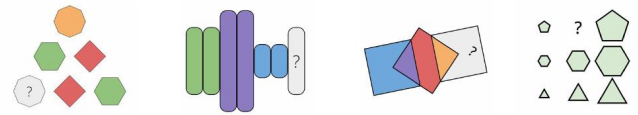
Gradient Tasks

Example: *Color Gradient Perception & Application*
(color gradient)

Prompt Example: *The part denoted by the question mark should be completely filled with the correct color.*

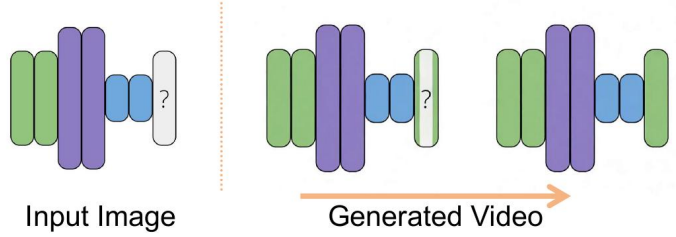


Compositionality Tasks



Example: *Rectangle Height Color Matching*
(compositionality of color and shape)

Prompt Example: *The part denoted by the question mark should be completely filled with the correct color.*



10 类任务

由 PuzzleVQA 改编

三种任务类别

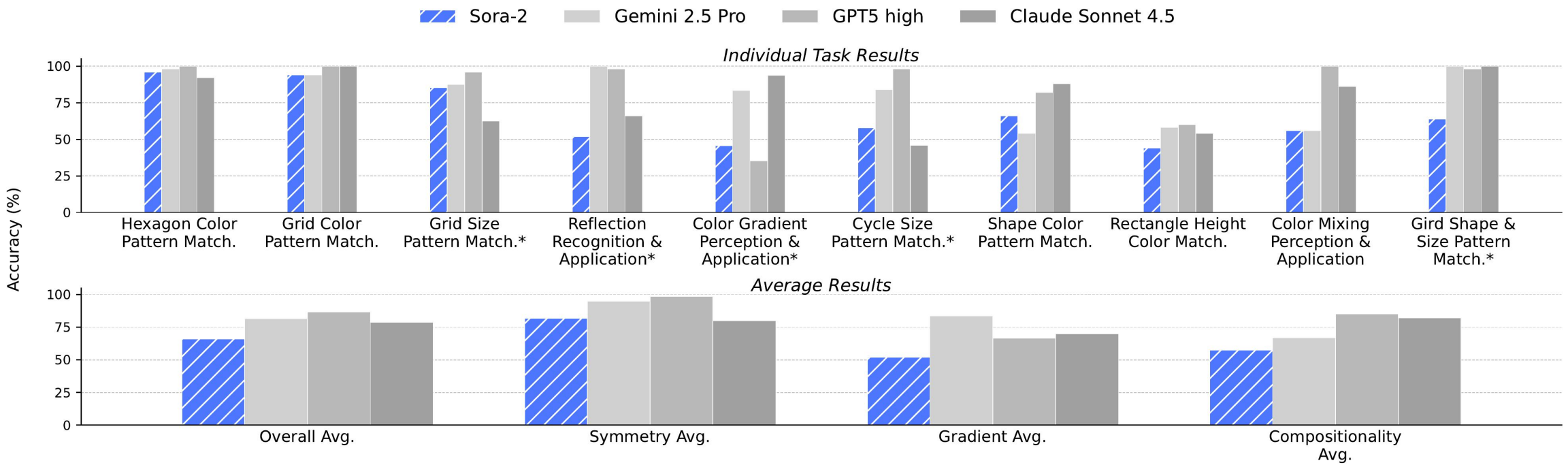
Symmetry / Gradient / Compositionality

生成式回答

Sora-2 不依赖选择项

Visual Puzzles：结果

Sora-2 具备一定视觉归纳推理能力



Symmetry 平均

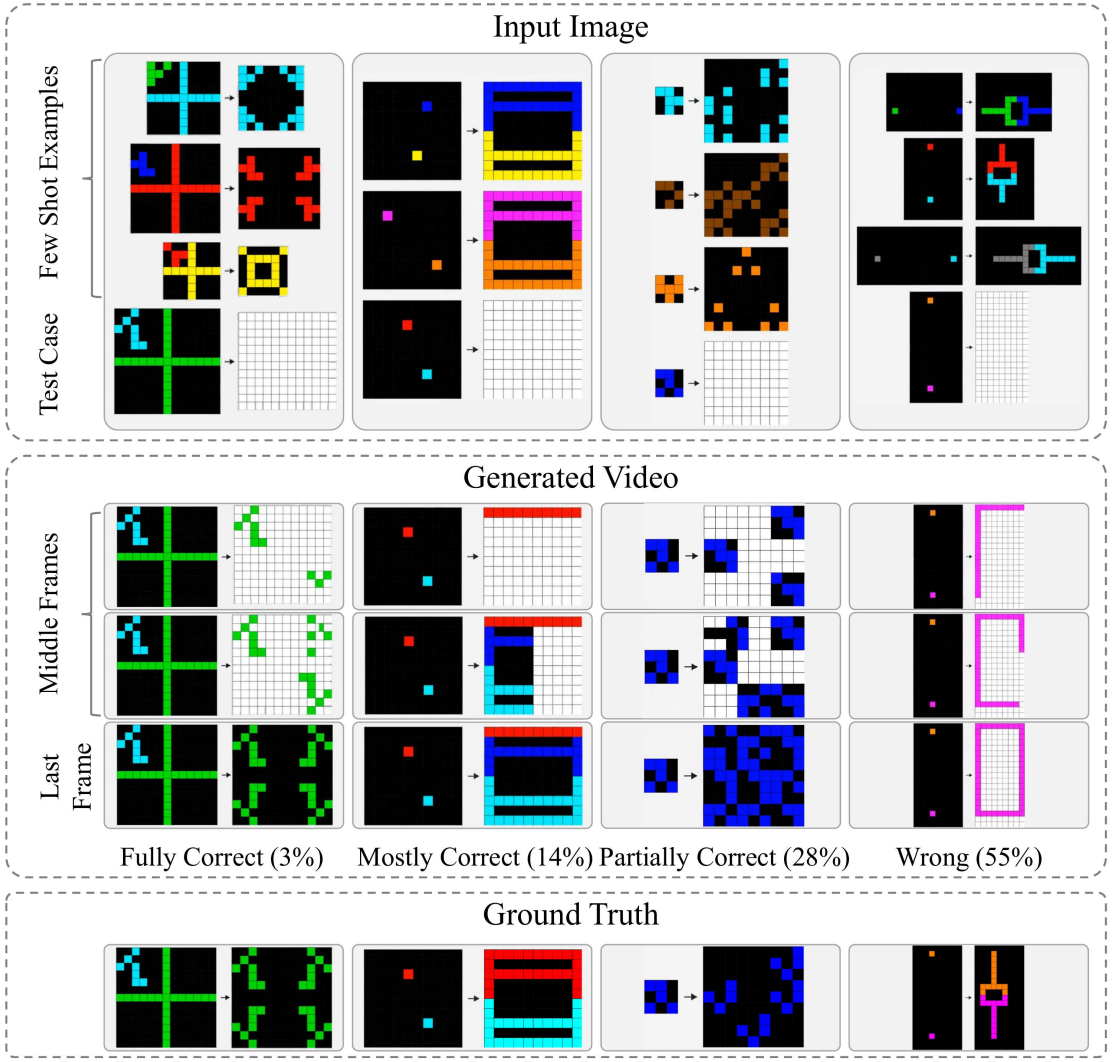
81.9%，超过 Claude 的 80.1%

Overall Average

66.2%，说明 Sora-2 可识别并应用视觉规律

ARC-AGI-2：抽象规则归纳

从少量样例推断网格变换规则（Few-shot Learning）



输入图片

若干 input-output 示例
+ 一个测试输入

输出视频

完成目标网格区域的内容

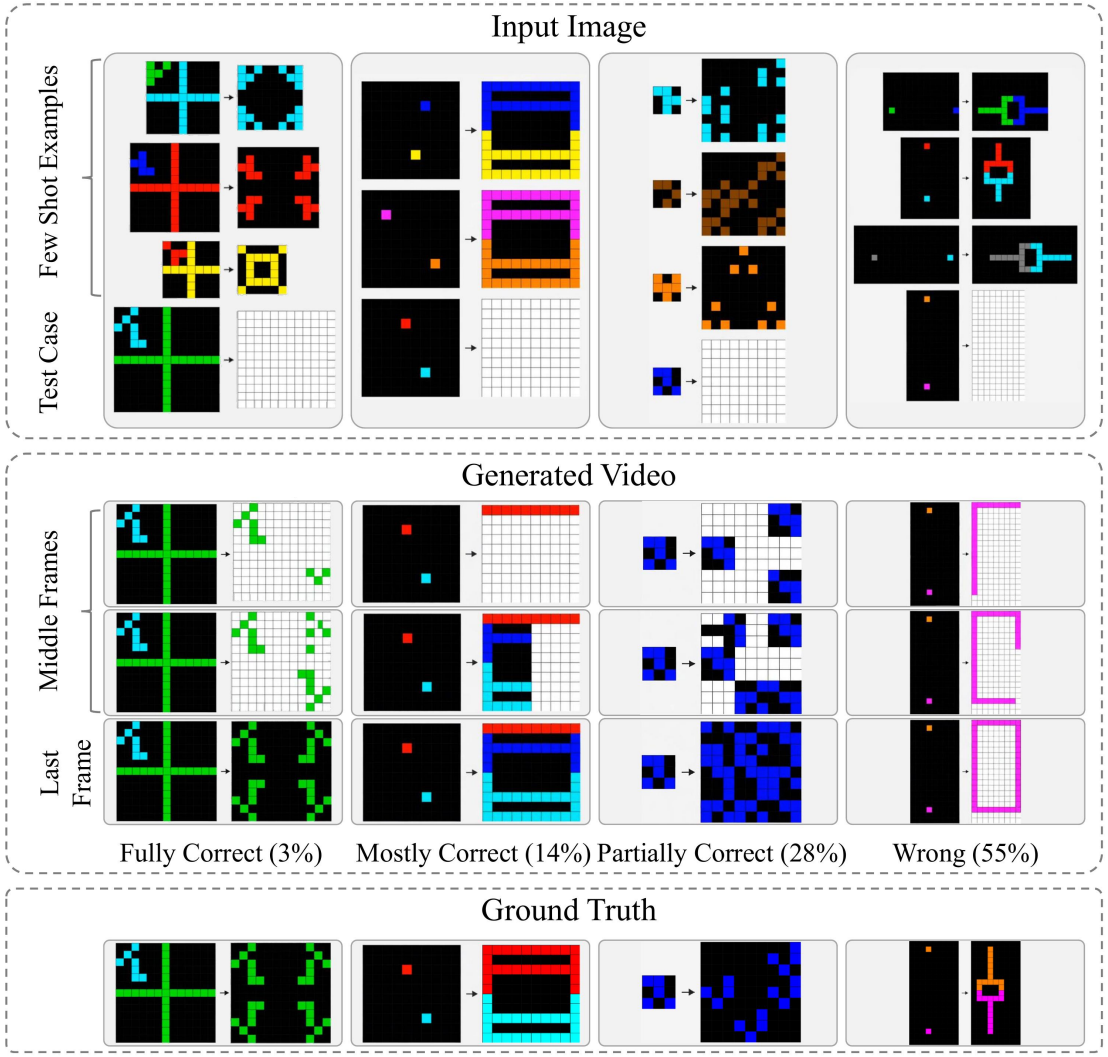
评估

比较视频最终帧与 ground truth

Sora-2 是 Few-shot Learner!

ARC-AGI-2：抽象规则归纳

匹敌 SOTA VLM



统一输入形式下（视觉输入）对比不同的模型：

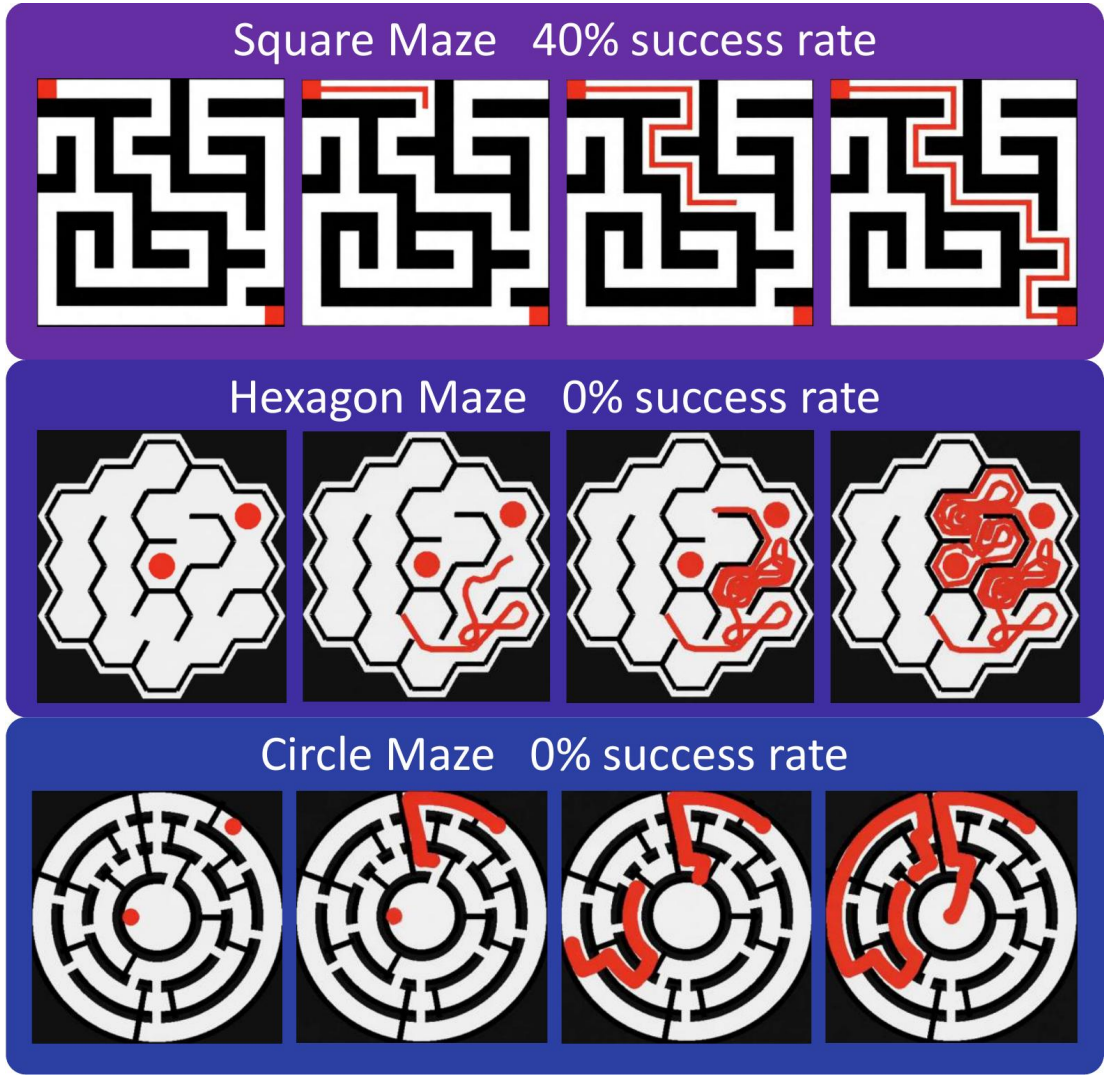
- SOTA VLM 也表现欠佳
- Sora-2 可与之匹敌
- Sora-2 还有 14% 是 “Mostly Correct”

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: Section 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

Mazes：空间规划与搜索

基于视频生成路径规划



任务要求

逐步规划形成内部路线

Square Maze

成功率 40%

Hexagon / Circle Maze

仍为 0%

而三个 VLM 在三种迷宫上得分均为 0!

Vision-Centric Reasoning

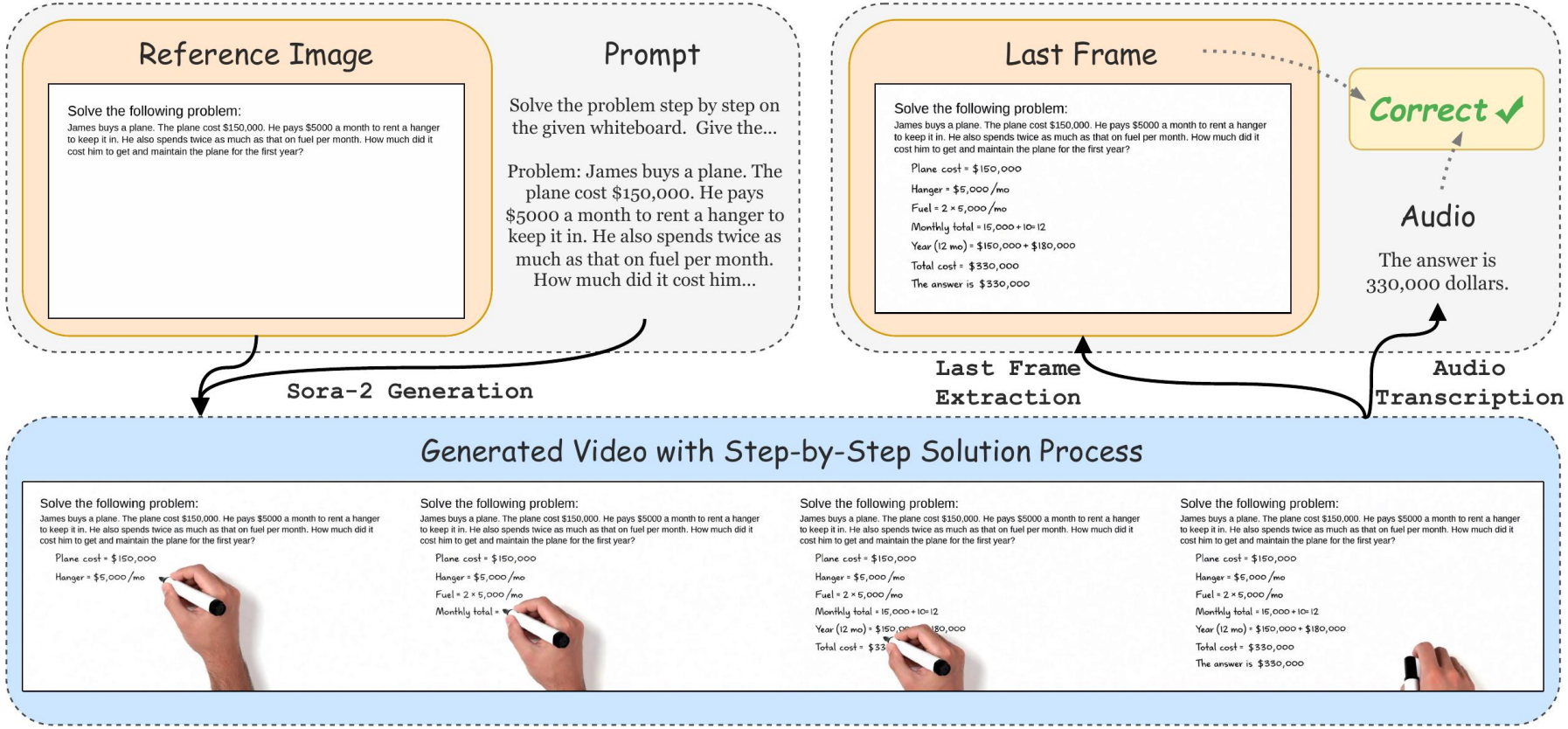
小结

Task	Sora-2	Gemini 2.5 Pro	GPT-5 high	Claude Sonnet 4.5
Eyeballing-Point	44.7	27.8	33.6	36.2
Eyeballing-Line	38.0	21.0	24.0	26.3
Eyeballing-Shape	34.5	34.5	32.5	50.5
Visual-Symmetry	81.9	94.9	98.5	80.1
Visual-Gradient	51.9	83.7	66.7	69.9
Visual-Comp.	57.5	67.0	85.0	82.0
ARC-AGI-2	1.3	1.9	0.5	5.3
Maze	13.3	0.0	0.0	0.0
Average	**40.4**	**41.3**	**42.6**	**43.8**

Sora-2 在视觉任务上总体可匹敌 SOTA VLM，展现出 Thinking with Video 的独特优势

Text-Centric Reasoning

视频生成模型也能处理文本推理题吗



输入

文本 prompt + 题目图像

输出

视频书写过程 + 音频答案

评估

Last Frame 与 Audio 分别评测

Text-Centric Reasoning

Sora-2 出人意料的文本推理表现

Dataset	Sora-2 Last Frame	Sora-2 Audio	Gemini 2.5 Pro	GPT-5 high	Claude Sonnet 4.5
			Text-Only Math Reasoning		
GSM8K	75.7	98.9	98.9	100.0	100.0
MATH-500	67.0	92.0	99.0	99.0	98.0
AIME24†	38.3	46.7	93.3	95.0	75.0
AIME25†	33.3	36.7	88.0	94.6	87.0
Average	53.6	68.6	94.8	97.2	90.0
			Text-Only General Knowledge Reasoning		
BBH	69.8	80.6	90.0	94.6	93.8
MMLU	69.1	67.3	87.7	86.0	89.5
MMLU-Pro	72.0	76.5	87.1	91.4	95.7
GPQA	51.5	57.6	86.4	85.7	83.4
SuperGPQA	53.2	44.5	71.1	68.3	69.0
Average	63.1	65.3	84.5	85.2	86.3
			Multimodal Math Reasoning		
MathVista	67.6	75.7	70.0	67.5	72.5
MathVision	44.9	46.7	63.3	71.6	58.7
Average	56.3	61.2	66.7	69.6	65.6
			Multimodal General Knowledge Reasoning		
MMBench	60.4	89.0	86.9	84.2	82.5
MMMU	38.3	69.2	79.0	77.0	82.0
Average	49.4	79.1	83.0	80.6	82.3
Overall Average	57.0	67.8	84.7	85.8	83.7

多个文本测试集上取得亮眼表现

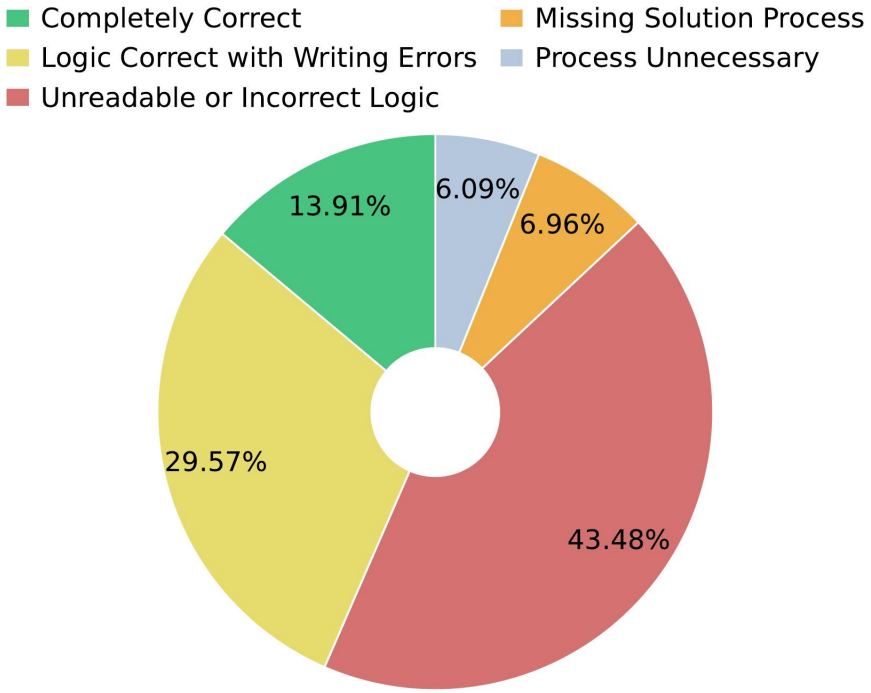
- 纯文本问题：GSM8K 上 98.9%，MATH 上 92.0%
- 多模态问题：MathVista 上 75.7%，MMMU 上 79.1%
- 较难的测试集上明显不如 SOTA VLM

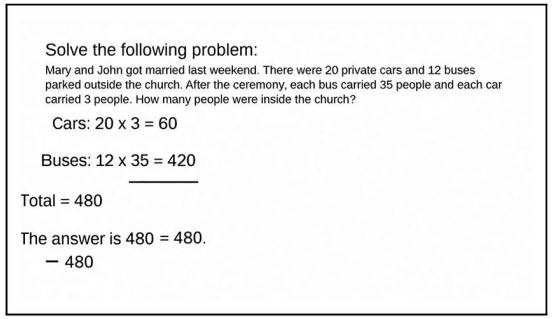
Audio 整体高于 Last Frame

说明答案常可被说出，但视觉书写更难稳定生成

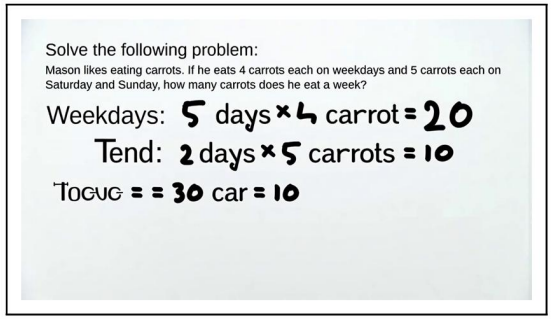
Text-Centric Reasoning

视频书写过程未必清晰可靠

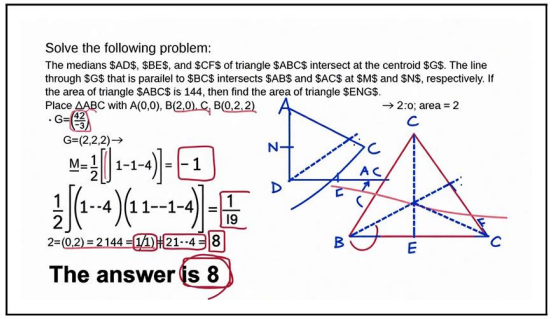




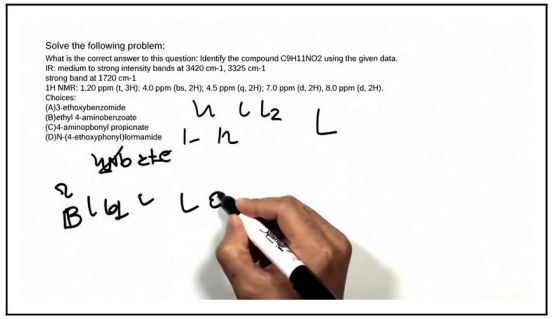
Completely Correct (13.91%)



Logic Correct with Writing Errors (29.57%)



Unreadable or Incorrect Logic (43.48%)



Missing Solution Process (6.96%)

模型会给出正确答案，但难以生成清晰、稳定和完全正确的推理步骤

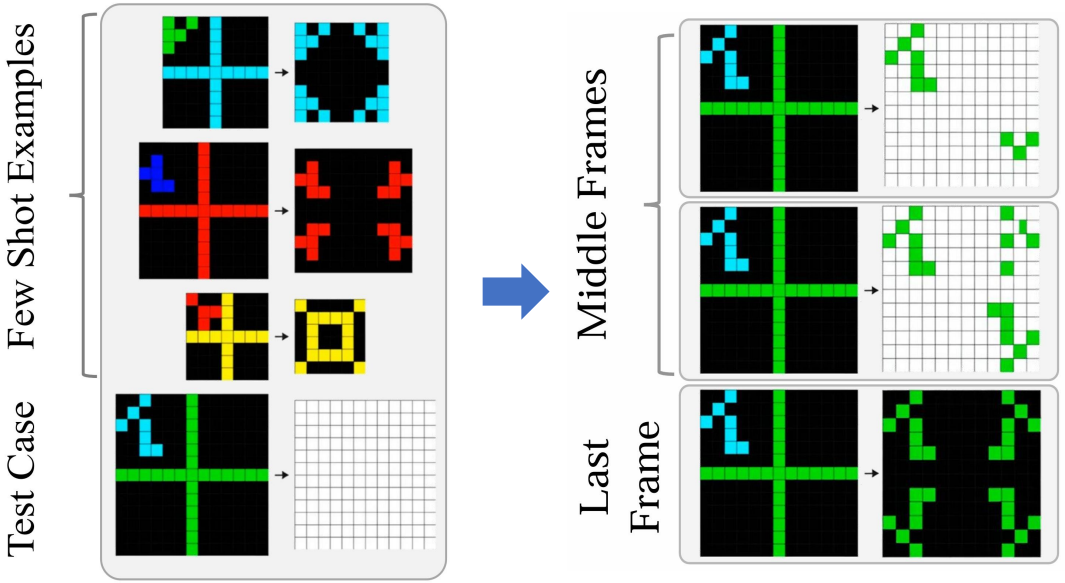
分析性实验 1

更多 Few-shot 示例能提升 ARC-AGI-2 上的表现

Few-Shot 使用全部示例，1-Shot 只给第一个示例

Accuracy Range	Few-Shot	1-Shot
0.00–0.35	743	788
0.35–0.65	127	117
0.65–1.00	130	95

Few-Shot 相比 1-Shot，更多测试样本有较高的 Pixel-Level 的准确率

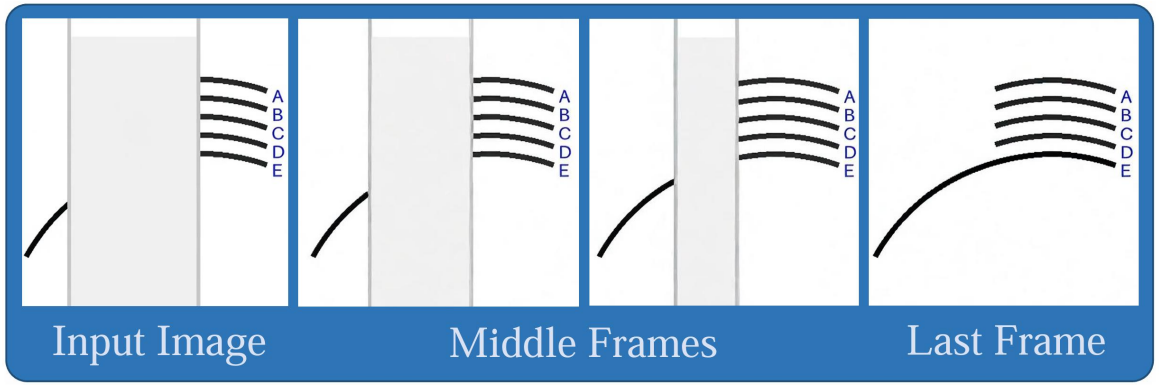


结论：视频生成模型的 In-context Learning 是一个值得探索的研究方向

分析性实验 2

Self-consistency 可进一步提升 Eyeballing Puzzle 上的表现

Last Frame	→	Major Frame	→	Major Frame + 5 Votes
56%		68%		90%



结论：视频生成推理的 test-time scaling 也值得探索

分析性实验 3

文本测试集上的亮眼表现是否只是因为测试集泄露?

- 改编测试数据（GSM8K 与 MATH），修改问题中的数值和表述
- 看 Sora-2 的表现是否下降

Dataset	Last Frame	Audio
GSM8K	75.7	98.9
GSM8K (Derived)	78.4	100.0
MATH-500	67.0	92.0
MATH-500 (Derived)	75.0	91.0

结论

性能没有随数值改写明显下降，因此是模型能力，而非测试集泄露

分析性实验 4

Sora-2 的文本推理表现是否来自于一个前置的 VLM 提示词改写模块?

- 无法控制 Sora-2 启用/禁止 prompt 改写
- 因此使用 **Wan 2.5** 进行实验（API 提供了参数来控制是否允许改写 prompt）

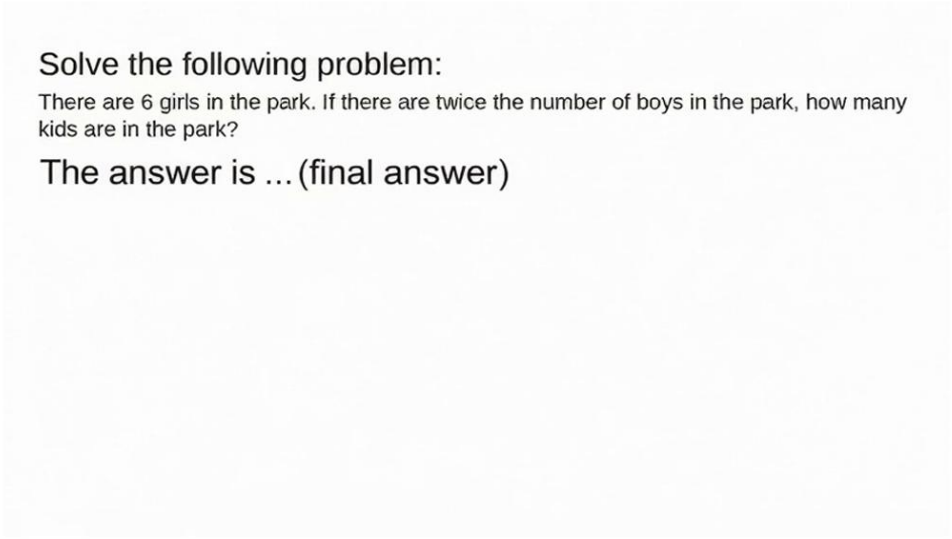
Dataset	Prompt Rewrite	Last Frame	Audio
GSM8K	✘	0.0	0.0
	✔	**78.4**	**31.9**
MMLU	✘	0.0	0.0
	✔	**74.1**	**50.00**
MMMU	✘	2.0	0.0
	✔	**47.0**	**14.0**

推测：Sora-2 的文本推理能力也可能来自内部 prompt rewriter

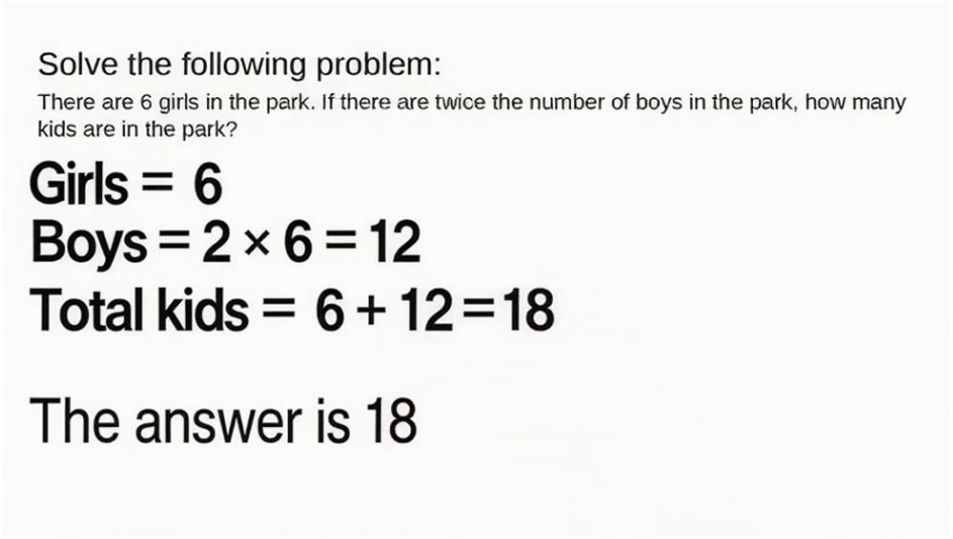
分析性实验 4

Sora-2 的文本推理表现是否来自于一个前置的 VLM 提示词改写模块?

- 无法控制 Sora-2 启用/禁止 prompt 改写
- 因此使用 **Wan 2.5** 进行实验（API 提供了参数来控制是否允许改写 prompt）



(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 (Appendix E.3.1).

推测：Sora-2 的文本推理能力也可能来自内部 prompt rewriter

总结

Sora-2 等视频模型具有独特多模态推理优势，并有望统一多模态生成与理解

视觉任务

可通过绘制与想象完成部分推理

文本任务

视频模型能在生成的视频中承载文本来进行推理

能力边界

文本书写过程不稳定，内部机制仍需进一步分析

Thinking with Video 是有潜力的统一多模态推理范式

未来工作

让视频模型真正学会推理

1. 更多模型和 SFT 训练

- ✓ 扩展更多开源视频生成模型评测
- ✓ 训练开源视频模型，如监督微调（SFT）

2. RLVR 强化训练

- ➢ Scale up 可验证视频生成推理数据，用 RLVR 增强 Thinking with Video 的能力和泛化性

3. 文本能力训练

- ➢ 将文本语料转为视频式训练数据
- ➢ 探索视频模型是否能内化文本语义和推理能力

目标：迈向更加强大，且可泛化的通用视频生成推理

https://thinking-with-video.github.io/#leaderboard
https://video-reason.com/

谢谢

Q&A



Paper: https://arxiv.org/abs/2511.04570

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

Website: https://thinking-with-video.github.io

Benchmark: https://huggingface.co/datasets/OpenMOSS-Team/VideoThinkBench