Evaluating Reasoning Coherence in Video Generative Models with Text and Visual Hints

Yu Qi[*†1], Xinyi Xu[*1], Ziyu Guo[*†‡2], Siyuan Ma[*2], Renrui Zhang[2], Xinyan Chen[2], Ruichuan An[3], Ruofan Xing[1], Jiayi Zhang[1], Haojie Huang[1], Pheng-Ann Heng[2], Jonathan Tremblay[4], and Lawson L.S. Wong[1]

[1] Northeastern University, USA
[2] The Chinese University of Hong Kong, China
[3] Peking University, Beijing, China
[4] NVIDIA, USA

Abstract. Video generative models show emerging reasoning behaviors. It is essential to ensure that generated events remain causally consistent across frames for reliable deployment, a property we define as ***reasoning coherence***. To bridge the gap in literature for missing reasoning coherence evaluation, we propose ***MME-CoF-Pro***, a comprehensive video reasoning benchmark to assess reasoning coherence in video models. Specifically, MME-CoF-Pro contains 303 samples across 16 categories, ranging from visual logical to scientific reasoning. It introduces Reasoning Score as evaluation metric for assessing process-level necessary intermediate reasoning steps, and includes three evaluation settings, (a) no hint, (b) text hint, and (c) visual hint, enabling a controlled investigation into the underlying mechanisms of reasoning hint guidance. Evaluation results in 7 open and closed-source video models reveals insights including: (1) Video generative models exhibit weak reasoning coherence, decoupled from generation quality. (2) Text hints boost apparent correctness but often cause inconsistency and hallucinated reasoning (3) Visual hints benefit structured perceptual tasks but struggle with fine-grained perception. We provide our project page along with data and code at: `https://video-reasoning-coherence.github.io/`.

Keywords: Benchmark and evaluation · Video reasoning · Video generative models

1 Introduction

Video generative models have rapidly progressed, producing high-fidelity outputs with realistic temporal dynamics. Emerging evidence [13, 48] suggests reasoning abilities such as planning and causal inference. It is now possible to prompt a generative video model to simulate an apple falling onto a jar (Fig. 1), predict the causal outcome of closing a window pane on a computer screen, or even perform zero-shot object identification from a single image. While these emerging abilities remain largely understudied by existing benchmarks, it is essential to ensure that generated events remain causally consistent across frames for reliable

∗ Equal contribution † Project Lead ‡ Corresponding Author

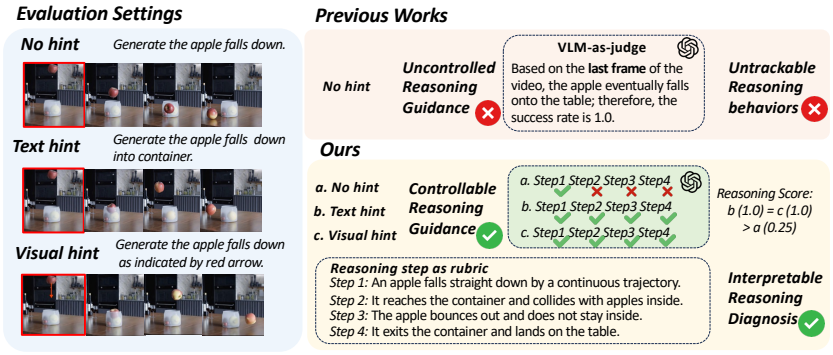


Fig. 1: *MME-CoF-Pro* evaluates video model capacity to model world behaviour. This evaluation is shown under the numbered multi steps section (shown in darker green), this matches with "reasoning steps as rubric" section (human generated to match expected behaviour). Using different prompting mechanisms, our benchmark evaluates video model behaviour at a fine grain. Red box indicates input image.

deployment. We formalize this gap as *reasoning coherence*: the extent to which generated events follow consistent and plausible cause–effect dynamics across frames.

Recent efforts such as V-ReasonBench [33] and Gen-ViRe [28] have taken steps toward evaluating reasoning in video generative models, introducing benchmarks across spatial, physical, and abstract dimensions. While these approaches move beyond surface-level fidelity, they primarily assess outcome correctness through last-frame verification. Concurrently, VIPER [25] and VBVR [45] introduced process-aware and large-scale verifiable evaluation of intermediate reasoning steps, but do not target the consistency of cause-and-effect chains specifically. In parallel, Morpheus [54] and VideoPhy [4] evaluate physical plausibility of individual phenomena but do not assess whether multi-step causal relationships remain coherent across the full generated sequence. While these directions are promising, none explicitly capture what we define as ***reasoning coherence***: the degree to which generated video events maintain consistent cause-and-effect relationships over time, spanning long-horizon dependencies, interaction reasoning, and the fidelity of intermediate transitions.

To bridge this gap, we introduce ***MME-CoF-Pro***, a benchmark for evaluating reasoning coherence in video generative models. In greater details, it covers 16 categories and 303 samples, as illustrated in Fig. 2. We introduce two key components for evaluating reasoning coherence, the reasoning score, and visual and text prompt-hints. First, the **Reasoning Score (RS)** provides process-level evaluation by assessing intermediate reasoning steps which are necessary in order to accomplish a task, allowing fine-grained analysis of how reasoning unfolds across frames. Each sample in MME-CoF-Pro is paired with a human-validated step-level rubric, scored automatically over k equidistant frames sampled from the generated video. Second, we design three controllable hint prompting settings: **(a) no hints** (standard setting), **(b) text hints** (Fig. 2), and **(c) vi-**

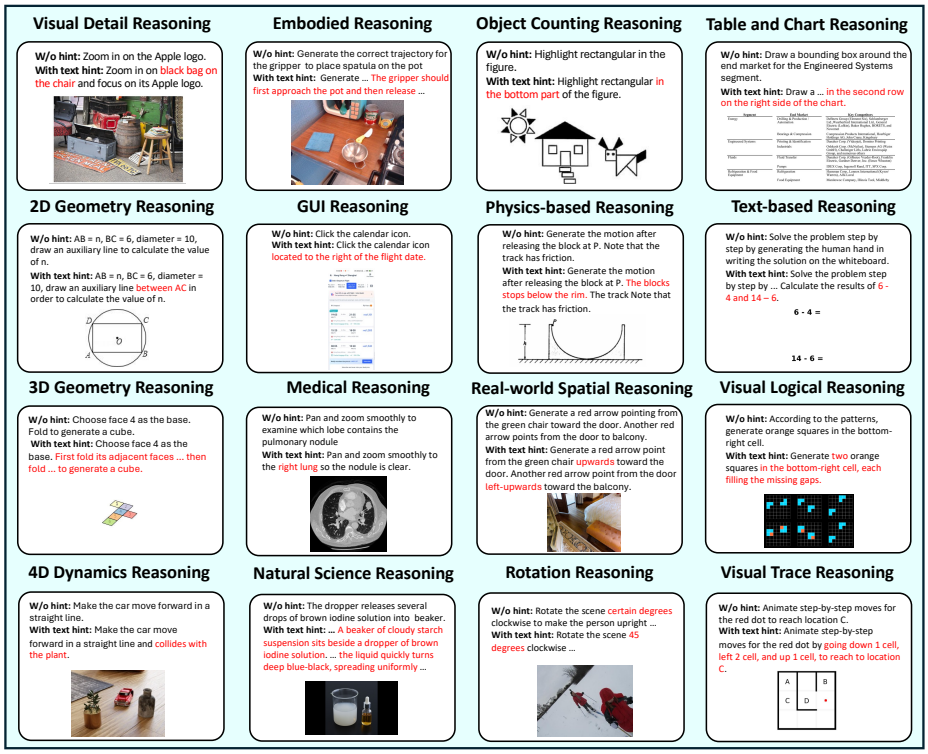


Fig. 2: ***MME-CoF-Pro*** consists of 303 samples from 16 basic reasoning categories, covering from *Visual Detail Reasoning* to *Visual Trace Reasoning*. Each sample is provided with no-hint and text-hint evaluation prompt. In text-hint setting, models are provided with additional textual details to facilitate reasoning.

sual hints (Fig. 3). The hint-based settings provide explicit reasoning scaffolds through visual or text guidance, whereas the no-hint setting requires models to reason without external assistance. As shown in Fig. 5, all remaining parts of instructions are kept identical across settings so any performance differences can be directly attributed to the presence of structured reasoning guidance.

Evaluation results of ***MME-CoF-Pro*** across 7 open and closed-source models reveal insights: **(1)** Video generative models exhibit generally weak reasoning coherence, and their reasoning performance shows no clear correlation with generation quality. **(2)** Explicit text hints may improve apparent correctness, but the gains are often superficial and can induce hallucinated reasoning, inflating scores without genuine understanding and leading to partial comprehension or overfitting to reasoning patterns. **(3)** Visual hints are more effective for structured and spatially guided tasks but are less reliable for fine-grained visual tasks. They also introduce hallucinations by being mistakenly rendered as part of the scene rather than purely as guidance. Our contributions can be summarized as follows:

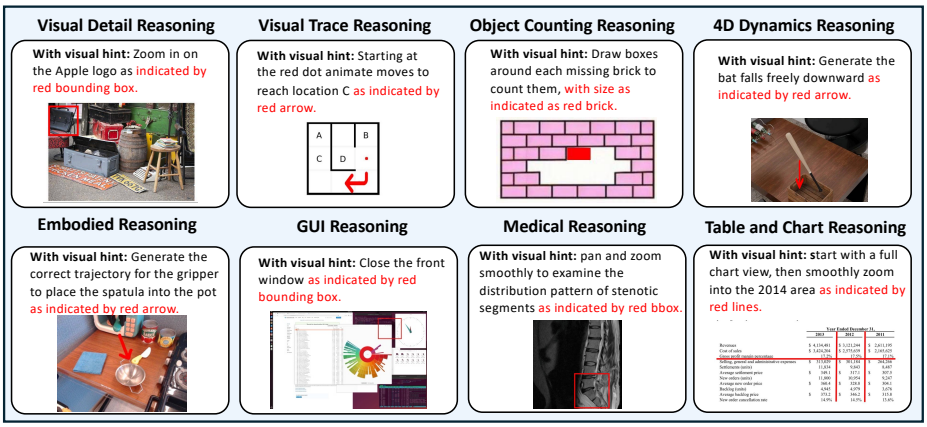


Fig. 3: Visual Hints in *MME-CoF-Pro*. For controlled comparison, we additionally add visual hints to 8 categories which requires visual reasoning capabilities, using bounding boxes and arrows to indicate positions, trajectories to facilitate reasoning.

- We introduce ***MME-CoF-Pro***, the first video reasoning benchmark that explicitly decouples reasoning hints evaluation which covers 303 samples across 16 categories (Fig. 4).
- We propose Reasoning Score, a process-level evaluation metric for evaluate intermediate reasoning steps instead of generation quality against human validated rubrics.
- Results on 7 open and closed-source video models uncover key insights including the performance of reasoning capabilities, limitations of explicit hints, the role of visual and text hints.

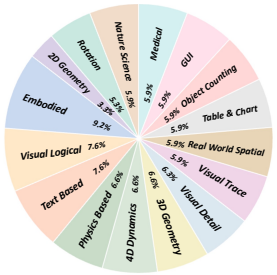
2 Related Works

2.1 Video Generative Models as Reasoning Models

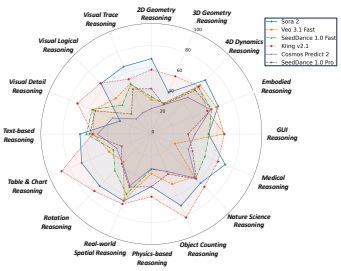
Recent advances in large reasoning models such as DeepSeek-R1 [14] have accelerated video reasoning. Prior work enhances MLLMs and VLMs [19, 32, 49] with causal reasoning [23, 34, 52], spatio-temporal modeling [6, 24, 27, 31, 53], and chain-of-thought [3, 12, 37, 55], using RL or supervised fine-tuning [21, 44, 47]. Recent large-scale video generative models [2, 7, 8, 13, 30, 38, 39, 43, 46] implicitly learn temporal dynamics, causal structures, and physical regularities through predictive video modeling. By forecasting future frames, diffusion-based and autoregressive frameworks [18, 50] capture object interactions, motion patterns, and long-horizon dependencies. Such temporal consistency and coherent scene evolution suggest that these models acquire structured world knowledge from large-scale video data, even without explicit reasoning supervision.

Statistic	Number
Total questions	673
- no hint questions	303
- text hint questions	303
- visual hint questions	67
Total samples	303
Total images	370
Total categories	16
Categories with visual hints	8
Max question words	191
Avg question words	81.6
Max reasoning step	10
Avg reasoning step	4.6
Avg extra word for text hint	17.4
Avg extra word for visual hint	5.2

(a) Key Statistics.



(b) Category Distribution.



(c) Evaluation Map.

Fig. 4: Statistics, Distribution and Evaluation Results of MME-CoF-Pro.

2.2 Reasoning Evaluation in Video Generative Models

Existing video reasoning benchmarks propose reasoning-related tasks, but they do not explicitly isolate and control reasoning as an independent evaluation variable in either the input design or the evaluation metrics [11, 16, 17, 25, 26, 28, 33, 40, 41, 51]. Prior work evaluates generation quality and creativity in T2V(Text-to-Video) models [22, 29], comprehensive reasoning tasks [15], maze-based spatial reasoning [34], synthetic dynamics prediction [33, 41], structured reasoning patterns [5], and scientific problem-solving [10, 20]. **(1)** Existing benchmarks embed reasoning guidance implicitly within task prompts rather than treating it as an explicit, controllable input. Without varying guidance independently, these benchmarks cannot measure how explicit reasoning instructions affect performance. **(2)** Existing evaluation metrics mainly focus on instruction alignment [15], visual consistency [15, 34], rule-based correctness [17], pass@k based final-frame correctness [5,33,41], spatial accuracy by hard coded rules [45], goal-conditioned success rate [35], process-consistency [25] but do not evaluate intermediate reasoning processes, limiting their ability to diagnose reasoning trajectory failures or providing targeted improvement suggestions.

In contrast to previous approaches, our proposed ***MME-CoF-Pro*** treats reasoning prompts as an explicit, controllable variable—injecting text and visual hints into the input while keeping the remaining instructions unchanged. By keeping all other task components identical, any performance difference can be causally attributed to reasoning capability and hint guidance. Finally, we introduce ***Reasoning Score***, a process-level evaluation metric for step-wise reasoning coherence (generated with humans), enabling interpretable diagnosis of where and why models fail.

3 The MME-CoF-Pro Benchmark

3.1 Overview of MME-CoF-Pro Benchmark.

In order to study reasoning behaviors of video generative models under controlled hint conditions, we propose ***MME-CoF-Pro***, the first benchmark that

treats reasoning guidance as an explicitly controllable variable and incorporates ***Reasoning Score*** as process-level evaluation metric for interpretable reasoning coherence diagnosis, as shown in Fig. 2, 3, and 4. *MME-CoF-Pro* contains 303 samples across 16 reasoning categories, spanning a broad range of reasoning capabilities, from perceptual tasks such as visual detail reasoning, to physically grounded and causal tasks including 4D Dynamics Reasoning. All samples include no-hint and text-hint settings. For a subset of 8 perceptually demanding categories, designated *MME-CoF-Pro-mini*, we additionally introduce a visual-hint setting (see Fig. 3). **(2)** We further annotate each sample with key reasoning steps as intermediate checkpoints (Fig. 6) and provide details in Section 3.1, allowing fine-grained diagnosis of reasoning trajectories. These annotations define the evaluation criteria for the Reasoning Score (RS).

Evaluation settings with no hint, text hint, and visual hint As illustrated in Fig. 2 and 3, ***MME-CoF-Pro*** evaluates each sample under controlled settings that differ only in the type of reasoning guidance provided. The **(a) no-hint** setting serves as the standard baseline, where models must reason purely from the task instruction without any additional scaffolding. The **(b) text-hint** setting augments the prompt with explicit textual descriptions of key reasoning steps, providing structured guidance to facilitate inference. For the 8 visually demanding categories in ***MME-CoF-Pro-mini***, we provide an additional **(c) visual-hint** setting (Fig. 3) in which bounding boxes or directional arrows are drawn on the input image to direct attention to relevant regions or motion directions. Since every part of the instruction except the hint remains identical across conditions, performance differences isolate the effect of reasoning guidance, supporting controlled and causal analysis of how models reason under different guidance types. Each hint guidance also shares the same reasoning score rubric for fair comparison.

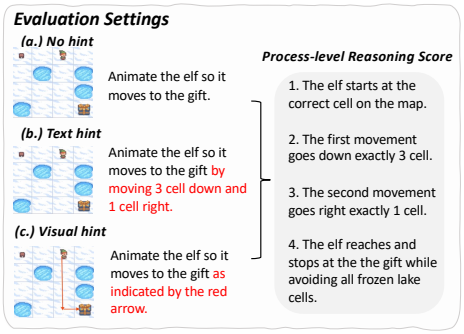


Fig. 5: Setting Comparisons. All three settings share the same evaluation metrics and identical instructions except for the hint components.

Reasoning Score (RS) over process-level key step. *MME-CoF-Pro* introduces a process-level evaluation framework that assesses reasoning coherence for each intermediate step which is necessary to accomplish a task. As illustrated in Fig. 6, each question is annotated as shown in Section 3.3 with a sequence of key reasoning steps, where each step represents a critical checkpoint for a correct video generation. Formally, given a generated video V and N annotated

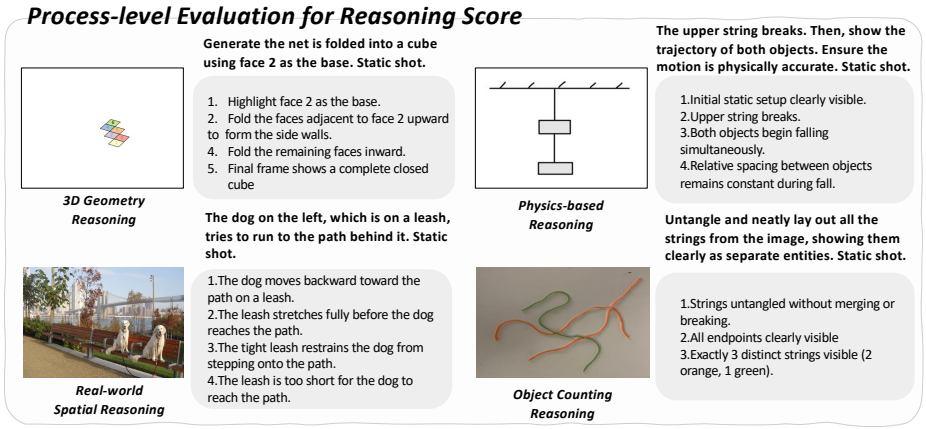


Fig. 6: Process-level Evaluation Metric for Reasoning Score in *MME-CoF-Pro.* Videos are evaluated against fine-grained reasoning steps, enabling interpretable assessment beyond final-answer accuracy.

reasoning steps $\{s_1, s_2, \ldots, s_N\}$, the Reasoning Score is computed as:

$$\text{RS} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[s_i \text{ is correctly completed}] \tag{1}$$

where each step s_i is evaluated independently by a judge model, such as Gemini-2.5-Flash [9]. This design allows fine-grained diagnosis of where along the reasoning chain a video generative model succeeds or fails, and supports more reliable cross-model comparison by capturing partial correctness rather than discarding it.

3.2 Category Distribution

The 16 categories in MME-CoF-Pro are organized into four groups, progressing from low-level perceptual understanding to high-level task-oriented reasoning, ensuring comprehensive coverage of the reasoning capabilities required for video generation.

Perceptual Reasoning covers fine-grained visual understanding, including *Visual Detail Reasoning* (identifying attributes by zooming into specific regions), *Rotation Reasoning* (interpreting scenes at non-standard orientations and restoring correct viewpoints), and *Object Counting Reasoning* (accurately enumerating objects of specific types within cluttered scenes).

Spatial and Structural Reasoning requires understanding of geometric and spatial relationships, spanning *Visual Trace Reasoning* (animating step-by-step navigation on mazes while avoiding obstacles), *Real World Spatial Reasoning* (inferring egocentric and allocentric spatial relationships in real-world environments), and *2D/3D Geometry Reasoning* (constructing auxiliary lines over planar figures and reasoning about 3D object structure across viewpoints).

Physical and Causal Reasoning evaluates grounded understanding of how

the world behaves over time, covering *Physics-Based Reasoning* (animating motion governed by physical laws such as gravity and collision), *4D Dynamics Reasoning* (predicting object interactions over time such as occlusion, collision, and containment), and *Nature Science Domain* (reasoning grounded in domain knowledge across chemistry, biology, and optics).
Task-oriented Reasoning covers reasoning embedded in real-world application contexts, including *Embodied Reasoning* (generating correct robotic manipulation trajectories), *GUI Reasoning* (identifying and interacting with interface elements in desktop or mobile screenshots), *Medical Reasoning* (analyzing clinical images such as CT scans and X-rays), *Table and Chart Reasoning* (locating specific data entries via zoom or bounding box operations), *Text-Based Reasoning* (mathematical problem solving and code execution tracing), and *Visual Logical Reasoning* (pattern completion and rule inference over structured visual grids and board games).

3.3 Benchmark Curation

Data Source. *MME-CoF-Pro* is curated from 27 existing real-world and synthetic benchmarks, including where each of the 16 reasoning categories is constructed by selectively sampling datasets that best represent the target reasoning skill (see Appendix for details). Real-world datasets provide ecological validity and grounded visual understanding, while synthetic benchmarks enable precise geometric, spatial, and procedural reasoning with controllable layouts and clear annotations. We respect all original licenses and usage terms and use only publicly available data. We provide more details in Appendix.

Data Annotation Pipeline. To ensure annotation reliability, each sample was annotated and reviewed by multiple human annotators. The reasoning steps were first drafted by Gemini-2.5-Pro [9] and subsequently refined through three rounds of manual verification and correction by domain experts. Images with visual hints are manually annotated by human annotators using Photopea, strictly preserving the original image resolution. During this process, annotators cross-checked each reasoning step to ensure that it was (1) necessary for the intended reasoning trajectory and (2) unambiguously verifiable from the generated video. Disagreements were resolved through discussion until consensus was reached. The annotated reasoning steps represent minimal necessary checkpoints required for successful task completion, which reduces ambiguity while preserving evaluation consistency.

4 Experiment and Analysis

After detailing the experimental setup including models, metrics, hint configurations, and human study design in Section 4.1, we validate the proposed Reasoning Score via human evaluation in Section 4.3. Section 4.2 assesses how well current video generative models reason and how reasoning relates to generation quality,

while Sections 4.4 and 4.5 explore whether reasoning hints enhance reasoning and how hint modality shapes the outcome.

4.1 Experiment Setup.

Models. We evaluated 7 closed and open-source models, Veo-3.1 [48], Veo-3.1-fast [48], Sora-2 [36], Seedance-1.0-pro [13], Seedance-1.0-fast [13], Kling-v2.1 [39] and Cosmos-Predict2-14B [1]. We use default configurations of these video models, with further inference configurations provided in Appendix.

Evaluation Protocols. Following recent video reasoning benchmarks, we adopt an VLM-as-a-judge protocol for automated evaluation, employing Gemini-2.5-Flash [9] as our judge model. During evaluation, each generated video is uniformly sampled at every 10th frame. Videos are 4 or 5 seconds long depending on the generative model, all at 24 fps. The reliability of judge model is validated via human study in Section 4.3, where RS achieves the highest Spearman correlation (0.61) with human scores.

Evaluation Metrics. We evaluate models from two perspectives: *Reasoning Score (RS)* and *Generation Quality (GQ)*. **(1)** For RS, Gemini uses the annotated process-level reasoning steps, as shown in Fig. 6, to identify whether each reasoning step is successfully achieved in the output video. **Reasoning Score is reported as percentages**, higher is better. **(2)** GQ is evaluated along five dimensions: Consistency Score (CS), temporal consistency, visual stability, hallucination, and physics grounding. CS measures cross-frame content coherence and the absence of abrupt content changes. Temporal consistency evaluates motion smoothness across frames. Visual stability assesses the steadiness of camera, objects, and scene composition. Hallucination penalizes unexpected or spurious visual artifacts. Physics grounding measures whether motions and interactions are physically plausible. We report the percentage value of CS and the average (Avg) across all five dimensions representing Generation Quality, and higher is better.

Evaluation Baselines. We evaluate three settings: *no hint*, *text hint*, and *visual hint*. The *no hint* and *text hint* settings are tested on the full *MME-CoF-Pro* (16 categories), where *no hint* serves as the baseline to measure gains from text hints. The *visual hint* setting is evaluated on *MME-CoF-Pro-mini*, a curated subset covering 8 perceptually intensive categories, with its corresponding *no hint* results as the baseline.

Human Study Setup. To validate that RS serves as an effective and relatively independent measure of video reasoning capability, we conduct a human study on 10 randomly selected generated videos with 10 human participants, each of whom scores the videos according to the annotated reasoning steps. Furthermore, we compare RS with two alternative metrics adopted from prior work:

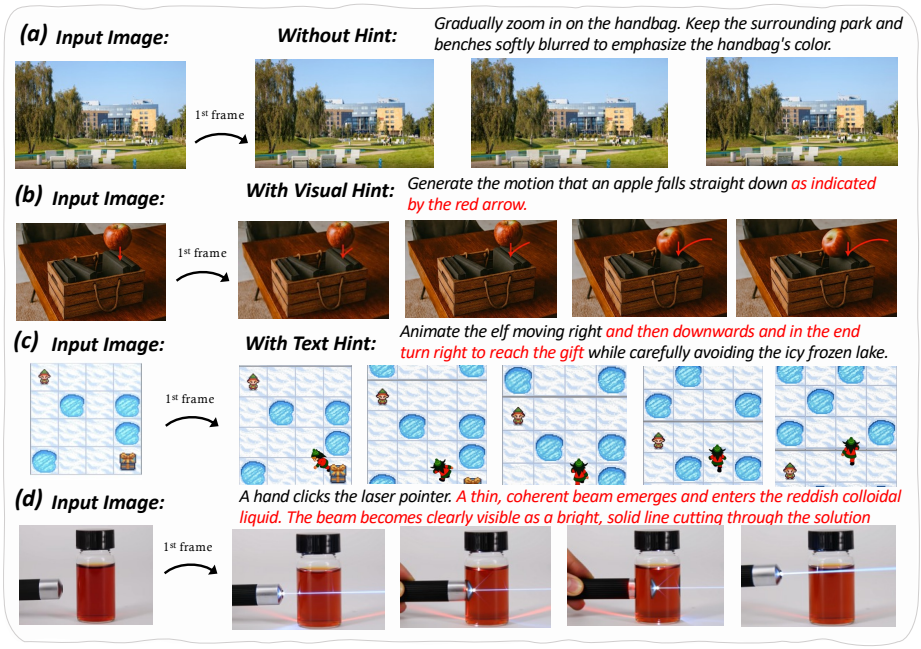


Fig. 7: Failure Examples of Video Generative Models. Subfigure (a) shows that the Kling model prioritizes perceptual fidelity (e.g., wind and forest dynamics) over the intended reasoning trajectory of gradual zoom-in. Subfigures (b) to (d) reveal hallucinations and inconsistencies arising under different hint conditions.

Instruction Alignment [15] and *Pass@5 last-frame correctness* [41]. The Spearsman correlation is computed between the average of human scores and each metric to assess their respective reliability in capturing reasoning performance.

4.2 Result Analysis of Reasoning Score

Overall models fall short in reasoning coherence. In Tab. 1, we report the RS under no-hint and text-hint settings. Model-level averages show clear stratification: frontier proprietary models lead, with Veo achieving the highest RS (56), followed by Sora (50), while Seedance, Cosmos, and Kling lag behind. Notably, even the best model only slightly exceeds 50, indicating that reasoning in video generation remains challenging.

Decoupling between generation quality and reasoning score. We observe a clear mismatch between overall generation quality (Avg) and RS, indicating that strong visual quality does not necessarily reflect strong reasoning ability. Some models achieve competitive Avg scores while exhibiting low RS, suggesting they generate visually coherent videos without correctly executing reasoning steps. An illustrative case is Kling, which achieves a high Avg score (65.1) but a

Table 1: Evaluation results of MME-CoF-Pro. We report the percentage of *Reasoning Score (RS)* and *Consistency Score (CS)* in **no-hint settings**, with text-hint performance shown in parentheses. Color indicates the value of change brought by **text-hint**: ↑ improvement and ↓ degradation. We additionally evaluate *Hallucination Score*, *Visual Stability*, and *Physics Grounding*, and present the overall average score with *Consistency Score* here. Details are provided in the Appendix.

Category	Metric	Closed-Source Models						Open-Source Models
		Veo3.1 [48]	Veo3.1-fast [48]	Sora2 [36]	Seedance 1.0-pro [13]	Seedance 1.0-fast [13]	Kling-v2.1 [39]	Cosmos Predict [1]
Visual Detail Reasoning	RS	61.5(+18.6)	65.7(+6.0)	40.3(+27.8)	19.8(+29.3)	26.1(+11.7)	2.0(-2.0)	30.2(+8.1)
	CS	51.2(0.0)	47.6(-1.8)	22.8(+13.3)	50.0(+7.6)	55.6(-13.3)	64.7(-16.6)	36.1(-10.6)
	Avg	56.8(-1.1)	56.7(-3.2)	30.7(+13.7)	54.8(+6.6)	58.1(-5.3)	72.3(-7.3)	37.9(-8.6)
Embodied Reasoning	RS	81.7(-0.2)	78.5(+0.8)	38.4(-3.9)	48.7(+6.8)	57.0(-6.4)	2.5(-1.4)	49.4(+20.1)
	CS	72.9(+1.1)	62.5(+8.2)	61.8(-1.1)	56.9(-3.3)	52.8(+0.1)	44.1(-3.8)	58.6(-6.6)
	Avg	76.1(+1.6)	63.4(+9.9)	65.4(-4.3)	58.9(-3.6)	58.8(-2.2)	54.8(-2.4)	57.9(-4.4)
Object Counting Reasoning	RS	51.7(+5.8)	55.8(-2.1)	57.1(+5.1)	35.4(+14.2)	31.2(+4.6)	11.8(-3.8)	16.6(-5.3)
	CS	47.2(-0.6)	46.7(+8.3)	69.4(-10.6)	36.5(-1.8)	30.0(+3.9)	75.3(-6.5)	36.1(+3.3)
	Avg	46.1(+3.4)	49.0(+7.1)	70.2(-2.9)	39.3(-2.2)	36.7(+3.2)	81.9(-4.0)	37.0(+11.3)
Table and Chart Reasoning	RS	49.4(+4.4)	52.3(-5.4)	61.3(+0.2)	33.7(-11.2)	29.9(+15.3)	33.5(-4.8)	23.8(+10.5)
	CS	31.1(+3.3)	38.3(-11.1)	72.2(-2.8)	25.7(-7.9)	33.3(+2.8)	85.7(-2.1)	33.9(+12.2)
	Avg	33.6(+3.3)	41.6(-8.4)	68.3(-2.8)	29.0(-9.3)	37.2(+1.8)	88.0(-5.6)	33.9(+5.2)
2D Geometry Reasoning	RS	29.2(-6.9)	22.2(-4.6)	54.2(-19.2)	22.5(-9.2)	9.2(+25.8)	0.0(+3.3)	9.2(-5.8)
	CS	28.0(-4.7)	27.8(+14.4)	64.0(+2.0)	32.0(+12.0)	22.0(+14.0)	46.0(+10.0)	15.0(+9.0)
	Avg	32.4(+0.5)	31.1(+12.2)	68.2(+2.6)	41.0(+11.8)	33.8(+3.8)	58.2(+12.4)	23.0(+4.4)
GUI Reasoning	RS	69.3(-10.1)	72.0(-4.6)	47.2(-3.1)	15.6(-0.4)	26.7(+1.1)	7.5(-1.8)	18.8(+5.7)
	CS	68.9(-6.1)	66.7(-2.2)	57.5(+8.8)	33.9(+2.8)	53.9(-2.2)	77.2(-15.6)	39.4(+17.2)
	Avg	69.7(-7.1)	65.6(+0.2)	51.8(+14.1)	33.2(+2.2)	52.3(-2.1)	66.0(-1.9)	39.6(+10.2)
Physics-based Reasoning	RS	55.9(+6.6)	65.8(-20.1)	62.2(+9.5)	53.2(-4.2)	45.5(-5.2)	18.9(-4.5)	25.9(+7.9)
	CS	30.5(-2.0)	36.5(-2.5)	51.0(+6.5)	48.5(-12.3)	34.5(+4.0)	45.4(+20.0)	30.5(-2.8)
	Avg	27.1(+2.5)	36.0(-4.4)	47.6(+7.6)	47.9(-7.9)	32.0(+5.0)	56.8(+6.3)	31.3(-1.3)
Text-based Reasoning	RS	62.3(+8.4)	62.0(+4.6)	75.4(+1.8)	35.6(-6.8)	44.9(-1.7)	20.7(-5.0)	29.1(-2.0)
	CS	49.6(+6.9)	49.6(+4.9)	55.6(-2.2)	49.6(-2.2)	54.8(-8.7)	34.8(+10.4)	43.5(-1.7)
	Avg	52.5(+9.1)	52.4(+9.9)	64.6(-5.0)	55.8(-3.7)	58.8(-8.1)	53.3(-0.8)	48.7(-3.3)
3D Geometry Reasoning	RS	43.9(+13.2)	41.7(+7.2)	37.7(+14.6)	36.4(-7.8)	48.5(-23.8)	13.4(+0.5)	22.1(+3.1)
	CS	27.2(+4.4)	27.2(+1.7)	42.8(-8.3)	28.2(-7.1)	22.8(-3.9)	50.6(+9.4)	26.7(+5.6)
	Avg	32.2(+0.1)	30.4(-2.8)	40.6(-2.9)	28.8(-4.4)	29.8(-5.1)	56.5(+4.5)	28.6(+9.2)
Medical Reasoning	RS	41.7(-0.2)	35.8(+3.2)	49.6(+10.0)	32.4(+10.1)	45.3(-6.4)	38.7(-2.5)	31.3(+13.0)
	CS	22.8(-1.1)	21.7(-3.3)	67.8(-7.2)	40.0(0.0)	51.2(-0.6)	54.4(+21.7)	37.8(+2.2)
	Avg	27.1(-2.6)	23.0(-2.2)	72.3(-10.4)	40.3(+1.0)	52.5(+3.1)	65.2(+6.0)	36.2(+6.2)
Real-world Spatial Reasoning	RS	54.0(-0.8)	58.7(+9.6)	35.0(+26.5)	30.6(-0.4)	26.2(+19.1)	9.4(-2.3)	41.4(+11.7)
	CS	57.8(-0.7)	63.3(+1.4)	62.6(-2.6)	53.2(-7.0)	60.0(-7.2)	58.4(+12.7)	52.6(+6.3)
	Avg	62.1(-0.4)	64.2(+1.8)	65.7(-0.3)	52.0(0.0)	59.0(-6.7)	68.6(+6.8)	50.6(+6.5)
Visual Logical Reasoning	RS	72.0(-7.4)	58.2(+8.9)	49.4(+12.3)	22.4(+11.2)	36.5(+3.2)	9.8(+4.6)	4.3(+49.3)
	CS	55.5(-5.5)	44.5(-1.9)	64.5(-7.6)	20.6(+2.6)	32.6(-6.5)	48.7(-1.7)	15.0(+14.6)
	Avg	55.5(-5.0)	45.7(+0.3)	64.5(-6.2)	25.6(+3.6)	37.1(-8.4)	65.1(0.0)	19.4(+10.4)
4D Dynamics Reasoning	RS	66.2(+12.4)	64.7(+11.3)	51.3(+6.9)	61.2(+2.7)	52.0(+9.1)	9.7(-0.9)	52.2(-3.4)
	CS	68.2(-1.2)	62.8(-7.2)	73.3(-13.9)	56.1(-11.1)	59.4(-2.2)	65.0(-15.6)	55.6(-11.1)
	Avg	65.4(+0.1)	62.4(-1.7)	69.0(-11.0)	56.0(-8.6)	60.0(-3.7)	60.8(-6.6)	46.8(-8.6)
Natural Science Reasoning	RS	49.6(+1.8)	55.3(+17.0)	60.3(-1.1)	46.0(-5.2)	49.6(+1.8)	11.3(+2.8)	51.0(+2.8)
	CS	57.3(-5.3)	51.1(+11.7)	66.7(-4.4)	53.8(+1.0)	57.3(-5.3)	62.8(-26.1)	57.5(-20.8)
	Avg	56.7(+0.7)	54.9(+9.0)	62.9(-0.3)	57.6(+0.6)	56.7(+0.7)	67.7(-14.9)	56.3(-15.3)
Rotation Reasoning	RS	60.2(-2.5)	56.0(+4.6)	30.1(+9.8)	41.8(+5.0)	35.8(-10.4)	11.9(+11.8)	25.0(+6.0)
	CS	44.4(+2.3)	34.4(+13.8)	71.2(-6.2)	40.0(+9.3)	47.5(-11.2)	79.3(-5.3)	41.2(-4.1)
	Avg	45.8(+2.1)	35.6(+9.6)	66.4(-0.6)	38.1(+10.9)	46.8(-12.5)	73.1(+2.5)	37.2(-5.5)
Visual Trace Reasoning	RS	45.5(+24.0)	49.8(+17.5)	49.7(+21.9)	35.7(+7.5)	36.0(+12.2)	19.5(-0.1)	27.9(+4.1)
	CS	47.8(+6.3)	43.3(+4.9)	63.9(-1.7)	38.9(+0.6)	44.4(-1.7)	40.0(+8.9)	20.6(+1.6)
	Avg	53.0(+5.7)	49.3(+1.8)	66.0(+2.3)	44.0(0.0)	48.7(-4.4)	53.7(+0.4)	20.7(+6.2)
Average	RS	55.9(+4.2)	55.9(+3.4)	49.9(+7.4)	35.7(+2.6)	37.5(+3.1)	13.8(-0.4)	28.6(+7.9)
	CS	47.5(-0.2)	45.3(+2.4)	60.5(-2.4)	41.5(-1.0)	44.5(-2.4)	58.3(-0.0)	37.5(+0.9)
	Avg	49.5(+0.8)	47.6(+2.4)	60.9(-0.4)	43.9(-0.2)	47.4(-2.6)	65.1(-0.3)	37.8(+1.4)

low RS (13.8), showing that strong visual fidelity can coexist with weak process-level reasoning. As shown in Fig. 7, the Kling model prioritizes the fidelity of wind dynamics over faithfully following the zoom-in instruction and searching for the handbag. This decoupling suggests that reasoning is a distinct capability, underscoring the need for dedicated reasoning-centric evaluation.

Takeaway 1

Video generative models exhibit generally weak reasoning ability, and their reasoning performance shows no clear correlation with generation quality.

4.3 Result Analysis of Human Study

Fig. 8: Human study spearman correlation map.

In Section 4.1, we validate whether Reasoning Score (RS) aligns better with human judgment than existing metrics. As shown in Fig. 8, RS achieves the highest Spearman correlation (0.61), outperforming *Instruction Alignment* [15] (0.17) and contrasting with *Pass@5 Correctness* [25](-0.41), which supports the claim that our proposed RS better captures human reasoning behavior and is an effective metric for reasoning coherence evaluation.

4.4 Result Analysis of Text Hints

Text hints consistently improve Reasoning Score but hurt consistency. Across 16 categories, the majority of models show positive RS changes when text hints are provided, for instance, Veo3.1 gains +4.5, Sora2 +7.6, and Cosmos Predict2-14B +6.7 at the overall average level, indicating that textual guidance helps models complete reasoning steps and improve task performance. However, this improvement comes at the cost of consistency: all of the 7 evaluated models experiencing CS degradation, Sora2 and Seedance 1.0-fast, suggesting that text hints broadly weaken temporal consistency and cross-video stability. Most strikingly, in 4D Dynamics Reasoning, all 7 models suffer CS drops under text hints (ranging from -1.2 to -15.6), raising the question of whether models are genuinely understanding the content or merely leveraging textual cues to produce superficially correct answers. As shown in Fig. 7(b)(c), these cases illustrate instruction-induced hallucination and inconsistency under textual hints. In the elf example, the model fails to ground the instruction in the original scene and instead 'forks' an extra elf to satisfy the motion directive, introducing a non-existent object and causing scene inconsistency. Overall, the model might overfits to textual hints and renders descriptions literally, prioritizing instruction completion over visual faithfulness and physical plausibility.

Table 2: Evaluation results (visual-hint). We report *Reasoning Score (RS)* and *Consistency Score (CS)* in **no-hint settings** for items that have visual-hint, with **visual-hint** change in parentheses. Color: ↑ improvement , ↓ degradation . Only categories with visual-hint data are shown.

Category	Metric	Closed-Source Models						Open-Source Models
		Veo3.1 [48]	Veo3.1-fast [48]	Sora2 [36]	Seedance 1.0-pro [13]	Seedance 1.0-fast [13]	Kling-v2.1 [39]	Cosmos Predict [1]
4D Dynamics Reasoning	RS	51.7(-0.4)	70.5(+8.5)	62.0(-3.9)	67.4(-25.8)	51.7(-6.6)	9.0(-5.6)	50.8(-7.0)
	CS	63.3(-15.6)	63.3(+2.2)	74.4(-12.2)	53.3(-4.4)	63.3(-21.1)	60.0(+3.3)	56.7(-2.2)
	Avg	62.4(-10.9)	62.2(+5.1)	71.6(-17.1)	55.3(-5.8)	62.4(-23.8)	61.8(-1.3)	52.0(+0.7)
Embodied Reasoning	RS	60.8(+22.9)	91.2(-17.5)	68.8(-8.3)	50.4(+8.3)	63.8(+12.4)	9.2(-2.9)	22.4(+20.5)
	CS	60.0(+17.5)	80.0(-3.8)	70.0(+5.0)	51.2(+8.8)	57.1(+17.1)	23.8(+2.5)	55.7(+12.9)
	Avg	61.5(+12.0)	81.0(-4.2)	69.2(+2.5)	58.8(+5.8)	60.3(+18.3)	42.0(+2.0)	56.3(+13.1)
GUI Reasoning	RS	27.5(+17.5)	82.5(+5.4)	48.3(+14.2)	25.0(+20.0)	27.5(+17.5)	2.1(0.0)	12.5(0.0)
	CS	50.0(+1.2)	60.0(-3.8)	40.0(+7.5)	35.0(-2.5)	50.0(+1.2)	71.2(+8.8)	35.0(+13.8)
	Avg	49.8(+1.8)	61.0(-6.8)	38.2(+13.0)	33.8(-6.0)	49.8(+1.8)	64.0(+9.5)	33.2(+10.8)
Medical Reasoning	RS	35.7(+6.4)	31.9(0.0)	37.9(+7.7)	21.2(+21.5)	35.7(-6.4)	46.4(+10.0)	23.1(+10.6)
	CS	48.6(-12.9)	27.5(+5.0)	47.5(+17.5)	37.5(+23.8)	48.6(-2.9)	65.7(0.0)	31.2(-12.5)
	Avg	48.9(-4.9)	29.2(+4.0)	56.8(+7.5)	40.0(+22.5)	48.9(+1.4)	72.0(-1.4)	26.0(+1.2)
Object Counting Reasoning	RS	40.0(+15.8)	60.8(-9.2)	76.7(+0.8)	36.7(-1.7)	37.5(+2.9)	9.2(+10.0)	26.7(-11.1)
	CS	40.0(-2.5)	58.8(-6.2)	75.0(-13.8)	48.8(-3.8)	28.8(+3.8)	70.0(+10.0)	35.0(+15.0)
	Avg	46.5(-6.5)	58.0(-6.8)	76.2(-13.5)	50.0(-0.8)	36.5(+3.0)	82.0(-0.5)	35.7(+16.0)
Table and Chart Reasoning	RS	26.7(+31.2)	52.7(-7.3)	56.7(-2.9)	31.7(-9.3)	26.7(+21.9)	34.8(+17.1)	9.5(+26.9)
	CS	33.8(+8.8)	45.0(-16.2)	61.2(-7.5)	28.6(-5.7)	33.8(+8.8)	84.3(+5.7)	17.1(+15.7)
	Avg	34.0(+6.0)	48.2(-18.0)	55.5(+7.8)	32.3(-6.3)	34.0(+4.5)	88.6(-8.0)	18.9(+13.7)
Visual Detail Reasoning	RS	27.8(-13.0)	45.8(+2.8)	37.6(+17.0)	22.6(-5.2)	27.8(-13.0)	4.2(-4.2)	12.5(+6.2)
	CS	54.4(-14.4)	56.7(-1.7)	20.0(+25.6)	47.8(-10.0)	54.4(-13.3)	68.8(-8.8)	42.5(+1.2)
	Avg	56.4(-10.4)	62.7(-2.3)	23.6(+28.0)	54.0(-15.3)	56.4(-8.7)	77.0(-10.8)	41.8(+4.0)
Visual Trace Reasoning	RS	23.7(+6.1)	35.9(-2.8)	40.0(+10.0)	29.4(-2.8)	23.7(+15.5)	16.4(0.0)	15.3(+2.1)
	CS	40.0(+2.2)	27.8(+7.8)	61.1(-5.6)	32.2(-6.7)	40.0(+7.8)	27.8(+15.6)	12.5(+17.5)
	Avg	42.0(+2.0)	34.4(+8.2)	59.6(-2.7)	40.0(-10.2)	42.0(+8.7)	52.2(+3.8)	14.5(+18.2)
Average	RS	36.9(+10.8)	58.9(-2.5)	53.5(+4.3)	35.5(+0.6)	36.8(+5.5)	16.4(+3.1)	21.6(+6.0)
	CS	48.8(-2.0)	52.4(-2.1)	56.2(+2.1)	41.8(-0.1)	47.0(+0.2)	58.9(+4.6)	35.7(+7.7)
	Avg	50.2(-1.4)	54.6(-2.6)	56.3(+3.2)	45.5(-2.0)	48.8(+0.6)	67.4(-0.8)	34.8(+9.7)

Takeaway 2

Text hints generally improve Reasoning Score, but consistently degrade Consistency Score and cause hallucination, suggesting that explicit guidance may shift model attention rather than enhance genuine understanding.

4.5 Result Analysis of Visual Hints

Visual hints do not consistently improve performance and can even degrade results on fine-grained visual tasks such as Visual Detail and Object Counting Reasoning. As shown in Tab. 2, several models exhibit notable drops in both RS and CS under visual-hint settings for detail-sensitive categories. For example, in Visual Detail Reasoning, Veo3.1 shows a 13.0 drop in RS and 14.4 in CS, while Seedance-1.0-fast drops by 13.0 in RS and 13.3 in CS. We hypothesize that visual hints add salient cues that bias the model toward coarse hint-following rather than precise visual grounding, causing it to overlook subtle details and thus degrade consistency and reasoning accuracy. In contrast, visual hints are more beneficial for structured or spatially guided tasks such as Embodied and GUI Reasoning, where directional cues better align with the reasoning process.

As shown in Fig. 7, visual hints can also induce hallucination and inconsistency, where the hint itself becomes part of the generated content. In many cases, the arrow used to indicate motion is implicitly rendered or transformed during generation (e.g., in Fig. 7(b) the arrow turning into a curved trajectory), suggesting that the model treats the visual hint as a drawable object rather than a guiding cue. We hypothesize that this behavior stems from training data distribution biases, where annotated visual markers (e.g., arrows, highlights) frequently co-occur with edited or synthetic content, leading the model to reproduce the hint as part of the scene instead of using it solely for guidance.

Takeaway 3

Visual hints are more effective for structured and spatially guided tasks but are less reliable for fine-grained visual tasks. They also introduce hallucinations by being mistakenly rendered as part of the scene.

4.6 Do More Hints Always Improve Reasoning Coherence? A Case Study

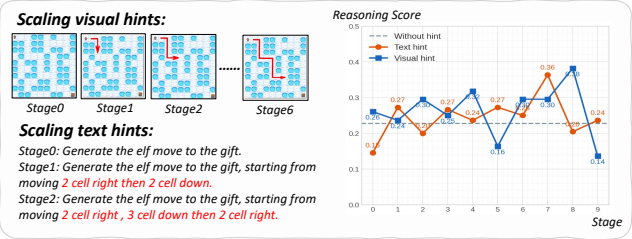


Fig. 9: Scaling case study on Frozen Lake.

A natural question arising from our hint analysis is whether providing more hints monotonically improves reasoning performance. To investigate this, we conduct a preliminary scaling experiment on the Frozen Lake [42] using Sora2, progressively increasing the number of text and visual hints and measuring reasoning score accordingly.

More specifically, as shown in the left part in Fig. 9, we define the scaling of stage by providing more visual (red arrows for potential directions) and correspondingly, more text instructions. For each setting, we roll out 10 times and report the average reasoning score. Results on the right suggest both text and visual hints generally yield reasoning scores above the no-hint baseline (0.23), suggesting that hint guidance is broadly beneficial. However, neither modality exhibits a monotonic improvement, both curves display substantial fluctuation across stages with no clear upward trend, indicating that current models cannot reliably leverage increasingly detailed hint information in a cumulative manner. This instability suggests that simply increasing hint information is insufficient to guarantee improved reasoning performance. This points to an open research direction: developing models capable of stably grounding multi-step hints into coherent reasoning trajectories.

5 Conclusion

In this work, we introduce MME-CoF-Pro, the first benchmark for reasoning coherence evaluation in video generative models. We propose *Reasoning Score* as process-level evaluation metric, and use three hint settings: *no-hint*, *text-hint*, and *visual-hint* to further discover hint-guided performance. Our experiments across seven state-of-the-art models reveal that reasoning ability remains limited and is largely decoupled from generation quality. We further show that hints reshape model behavior in modality-specific ways: text hints generally improve step completion but often reduce consistency and induce hallucinations. Visual hints tend to have task-dependent effects, they are helpful for structured and spatial reasoning, but less reliable for fine-grained perception and sometimes mistakenly incorporated into the scene itself. These findings suggest that current models tend to follow hints rather than truly ground them, highlighting the need for video generative models with stronger visual grounding, instruction understanding abilities and anti-hallucination mechanisms.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)
3. Arnab, A., Iscen, A., Caron, M., Fathi, A., Schmid, C.: Temporal chain of thought: Long-video understanding by thinking in frames. arXiv preprint arXiv:2507.02001 (2025)
4. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024)
5. Chen, H.H., Lan, D., Shu, W.J., Liu, Q., Wang, Z., Chen, S., Cheng, W., Chen, K., Zhang, H., Zhang, Z., et al.: Tivibench: Benchmarking think-in-video reasoning for video generative models. arXiv preprint arXiv:2511.13704 (2025)
6. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
7. Chi, X., Jia, P., Fan, C.K., Ju, X., Mi, W., Zhang, K., Qin, Z., Tian, W., Ge, K., Jia, Y., et al.: Wow: Scaling embodied omni-world model for generalizable manipulation simulation
8. Chi, X., Jia, P., Fan, C.K., Ju, X., Mi, W., Zhang, K., Qin, Z., Tian, W., Ge, K., Li, H., et al.: Wow: Towards a world omniscient world model through embodied interaction. arXiv preprint arXiv:2509.22642 (2025)
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
10. Deng, A., Yang, T., Yu, S., Spencer, L., Bansal, M., Chen, C., Yeung-Levy, S., Wang, X.: Scivideobench: Benchmarking scientific video reasoning in large multimodal models. arXiv preprint arXiv:2510.08559 (2025)
11. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27713–27724 (2025)
12. Fei, H., Wu, S., Ji, W., Zhang, H., Zhang, M., Lee, M.L., Hsu, W.: Video-of-thought: Step-by-step video reasoning from perception to cognition. arXiv preprint arXiv:2501.03230 (2024)
13. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
14. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
15. Guo, Z., Chen, X., Zhang, R., An, R., Qi, Y., Jiang, D., Li, X., Zhang, M., Li, H., Heng, P.A.: Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. arXiv preprint arXiv:2510.26802 (2025)
16. Haodong Chen, H., Lan, D., Shu, W.J., Liu, Q., Wang, Z., Chen, S., Cheng, W., Chen, K., Zhang, H., Zhang, Z., et al.: Tivibench: Benchmarking think-in-video reasoning for video generative models. arXiv e-prints pp. arXiv–2511 (2025)

17. He, X., Fan, Z., Li, H., Zhuo, F., Xu, H., Cheng, S., Weng, D., Liu, H., Ye, C., Wu, B.: Ruler-bench: Probing rule-based reasoning abilities of next-level video generation models for vision foundation intelligence. arXiv preprint arXiv:2512.02622 (2025)
18. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
19. Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., et al.: Cogvlm2: Visual language models for image and video understanding. arXiv preprint arXiv:2408.16500 (2024)
20. Hu, M., Chen, T., Zou, Y., Lei, Y., Chen, Q., Li, M., Mu, Y., Zhang, H., Shao, W., Luo, P.: Text2world: Benchmarking large language models for symbolic world model generation. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 26043–26066 (2025)
21. Hu, Y., Song, Z., Feng, N., Luo, Y., Yu, J., Chen, Y.P.P., Yang, W.: Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29108–29117 (2025)
22. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
23. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025)
24. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: European conference on computer vision. pp. 237–255. Springer (2024)
25. Li, Y., Gu, Y., Min, Y., Liu, Z., Du, Y., Zhou, K., Yang, M., Zhao, W.X., Qiu, M.: Viper: Process-aware evaluation for generative video reasoning. arXiv preprint arXiv:2512.24952 (2025)
26. Liang, A., Kong, L., Yan, T., Liu, H., Yang, W., Huang, Z., Yin, W., Zuo, J., Hu, Y., Zhu, D., et al.: Worldlens: Full-spectrum evaluations of driving world models in real world. arXiv preprint arXiv:2512.10958 (2025)
27. Liao, R., Erler, M., Wang, H., Zhai, G., Zhang, G., Ma, Y., Tresp, V.: Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 6577–6602 (2024)
28. Liu, X., Xu, Z., Li, M., Wang, K., Lee, Y.J., Shang, Y.: Can world simulators reason? gen-vire: A generative visual reasoning benchmark. arXiv preprint arXiv:2511.13853 (2025)
29. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22139–22149 (2024)
30. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
31. Liu, Z., Xu, K., Su, B., Zou, X., Peng, Y., Zhou, J.: Stop: Integrated spatial-temporal dynamic prompting for video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13776–13786 (2025)

32. Lu, H., Jian, H., Poppe, R., Salah, A.A.: Enhancing video transformers for action understanding with vlm-aided training. arXiv preprint arXiv:2403.16128 (2024)
33. Luo, Y., Zhao, X., Lin, B., Zhu, L., Tang, L., Liu, Y., Chen, Y.C., Qian, S., Wang, X., You, Y.: V-reasonbench: Toward unified reasoning benchmark suite for video generation models. arXiv preprint arXiv:2511.16668 (2025)
34. Maaz, M., Rasheed, H., Khan, F.S., Khan, S.: Video-r2: Reinforcing consistent and grounded reasoning in multimodal language models. arXiv preprint arXiv:2511.23478 (2025)
35. Maes, L., Lidec, Q.L., Haramati, D., Massaudi, N., Scieur, D., LeCun, Y., Balestriero, R.: stable-worldmodel-v1: Reproducible world modeling research and evaluation. arXiv preprint arXiv:2602.08968 (2026)
36. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., et al.: Open-sora 2.0: Training a commercial-level video generation model in 200 k. arXiv preprint arXiv:2503.09642 (2025)
37. Rasheed, H., Zumri, M., Maaz, M., Yang, M.H., Khan, F.S., Khan, S.: Videocom: Interactive video reasoning via chain of manipulations. arXiv preprint arXiv:2511.23477 (2025)
38. Seedance, T., Chen, H., Chen, S., Chen, X., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Cheng, T., Cheng, X., et al.: Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv:2512.13507 (2025)
39. Team, K., Chen, J., Ding, Y., Fang, Z., Gai, K., Gao, Y., He, K., Hua, J., Jiang, B., Lao, M., et al.: Klingavatar 2.0 technical report. arXiv preprint arXiv:2512.13313 (2025)
40. Tian, J., Li, S., He, C., Wu, L., Tan, C.: Envision: Benchmarking unified understanding & generation for causal world process insights. arXiv preprint arXiv:2512.01816 (2025)
41. Tong, J., Mou, Y., Li, H., Li, M., Yang, Y., Zhang, M., Chen, Q., Liang, T., Hu, X., Zheng, Y., et al.: Thinking with video: Video generation as a promising multimodal reasoning paradigm. arXiv preprint arXiv:2511.04570 (2025)
42. Towers, M., Kwiatkowski, A., Terry, J., Balis, J.U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al.: Gymnasium: A standard interface for reinforcement learning environments. arXiv preprint arXiv:2407.17032 (2024)
43. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
44. Wang, G., Zhou, Y., He, Z., Lu, K., Feng, Y., Liu, Z., Wang, G.: Knowledge-guided pre-training and fine-tuning: Video representation learning for action recognition. Neurocomputing **571**, 127136 (2024)
45. Wang, M., Wang, R., Lin, J., Ji, R., Wiedemer, T., Gao, Q., Luo, D., Qian, Y., Huang, L., Hong, Z., et al.: A very big video reasoning suite. arXiv preprint arXiv:2602.20159 (2026)
46. Wang, X., Zhu, Z., Huang, G., Wang, B., Chen, X., Lu, J.: Worlddreamer: Towards general world models for video generation via predicting masked tokens. arXiv preprint arXiv:2401.09985 (2024)
47. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International conference on medical image computing and computer-assisted intervention. pp. 101–111. Springer (2023)
48. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)

49. Xu, H., Ghosh, G., Huang, P.Y., Arora, P., Aminzadeh, M., Feichtenhofer, C., Metze, F., Zettlemoyer, L.: Vlm: Task-agnostic video-language model pre-training for video understanding. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 4227–4239 (2021)
50. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
51. Yang, C., Wan, H., Peng, Y., Cheng, X., Yu, Z., Zhang, J., Yu, J., Yu, X., Zheng, X., Zhou, D., et al.: Reasoning via video: The first evaluation of video models' reasoning abilities through maze-solving tasks. arXiv preprint arXiv:2511.15065 (2025)
52. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. arXiv preprint arXiv:1910.01442 (2019)
53. Yuan, Y., Zhang, H., Li, W., Cheng, Z., Zhang, B., Li, L., Li, X., Zhao, D., Zhang, W., Zhuang, Y., et al.: Videorefer suite: Advancing spatial-temporal object understanding with video llm. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18970–18980 (2025)
54. Zhang, C., Cherniavskii, D., Tragoudaras, A., Vozikis, A., Nijdam, T., Prinzhorn, D.W., Bodracska, M., Sebe, N., Zadaianchuk, A., Gavves, E.: Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. arXiv preprint arXiv:2504.02918 (2025)
55. Zhang, S., Hao, X., Tang, Y., Zhang, L., Wang, P., Wang, Z., Ma, H., Zhang, S.: Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 12745–12752 (2025)