MENTAL SIMULATION FOR VISION-LANGUAGE MODELS ON INTUITIVE PHYSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) excel at interpreting what is visible, but can often struggle with future-dependent questions whose answers hinge on events that have not yet occurred. Such questions require anticipating how a scene will unfold—something humans do naturally through mental simulation. We propose a test-time approach that gives VLMs an analogous capability. To complement a VLM's internal text-based reasoning, we use a generative video model to produce multiple plausible future rollouts conditioned on the observed video. These "mental simulations" are fed back to the VLM as additional visual context before it answers. We evaluate this simulate-then-reason framework on three intuitive physics benchmarks: Physion, CLEVRER, and real-world Physics-IQ videos. Across a diverse set of open-source and proprietary VLMs, augmenting models with imagined futures can improve prediction accuracy over both observed-only baselines and strong test-time language reasoning baselines, suggesting that allocating test-time computation to explicit video-based mental simulation can enhance multimodal reasoning about the physical world. Video examples are available on our project page.

1 INTRODUCTION

Imagine watching the first domino in a long chain begin to tilt. Even before it falls, you can anticipate the cascade of collisions that will follow. This ability reflects mental simulation: imagining how the scene might unfold beyond what is currently observed. Rather than reasoning only about what is currently visible, humans routinely anticipate what comes next, considering multiple plausible ways the world might unfold, before committing to a decision or action. Cognitive science has long suggested that such predictive simulation is a fundamental component of human perception and decision-making Battaglia et al. (2013); Schacter et al. (2007).

Following this, a natural question is whether modern vision-language models (VLMs) possess a similar capability. Recent progress in VLMs DeepMind (2025); OpenAI (2025); Bai et al. (2025) has led to substantial improvements in image and video understanding, which can be further improved by allocating additional compute at inference time. In particular, techniques such as chain-of-thought (CoT) prompting Wei et al. (2022) and test-time scaling Muennighoff et al. (2025) use this additional computation to encourage models to explore intermediate reasoning trajectories before producing an answer. While these approaches can endow VLMs with the ability to simulate the future, such reasoning is typically done in language space.

For physical reasoning, however, one can wonder if language alone is always the most effective space to carry out the prediction in. Consider the domino example again: determining whether the domino will cause a chain reaction requires anticipating how object positions, motion, and interactions evolve through a sequence of intermediate states. These dynamics may be describable through text, but they are inherently visual in nature. Representing the intermediate states visually may therefore provide a more direct way to reason about how the scene evolves over time. This motivates a complementary form of test-time reasoning that explicitly simulates possible visual futures and uses them as additional evidence for prediction.

Inspired by the human ability for mental simulation, we propose to augment VLM reasoning with an explicit mechanism for imagining possible futures. Rather than relying only on a VLM to extrapolate future dynamics in language, we generate multiple plausible future rollouts conditioned on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

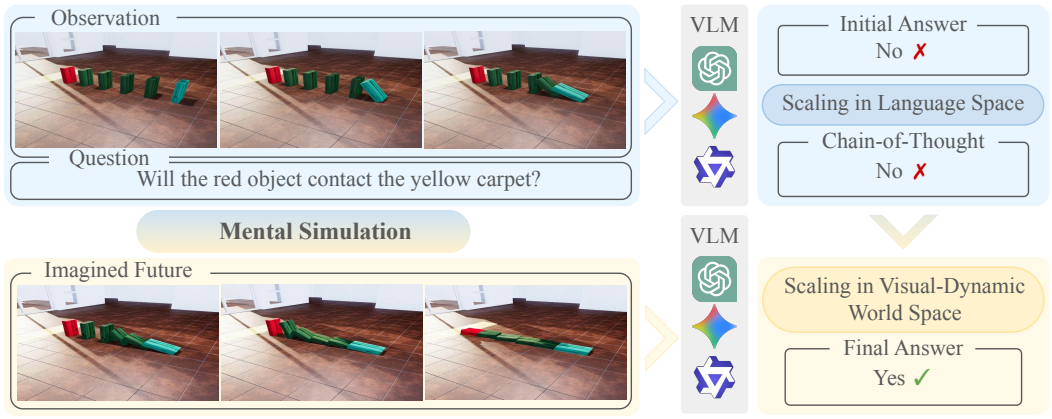


Figure 1: **Mental simulation enables future-aware reasoning.** When answers depend on what happens next, vision-language models may fail from observed frames alone. Generating imagined future rollouts can correct the prediction, illustrating test-time visual simulation for future-dependent prediction.

the observed context using a generative video model Wan et al. (2025); Chen et al. (2025); Song et al. (2025); Agarwal et al. (2026). These imagined futures are then provided to the VLM as additional visual evidence before answering the question. More broadly, we explore a complementary allocation of test-time compute by generating candidate future videos for a VLM to reason over.

We instantiate this simulate-then-reason paradigm on intuitive physics benchmarks where answering correctly requires predicting events beyond the observed video segment. Specifically, we evaluate on Physion Bear et al. (2021), CLEVRER Yi et al. (2020), and a repurposed subset of Physics-IQ Motamed et al. (2025) containing real-world physical interactions. Across a diverse set of open-source and proprietary VLMs DeepMind (2025); OpenAI (2026); Bai et al. (2025), augmenting models with imagined futures can improve prediction accuracy over both observed-only and test-time language reasoning baselines. These results suggest that explicit video-based mental simulation can provide useful additional evidence for future-dependent physical reasoning.

To summarize, we introduce visual mental simulation as a test-time mechanism for VLM reasoning about future physical events beyond an observed video. Our simulate-then-reason framework allocates test-time compute to generating multiple possible futures for the VLM to reason over. Across three intuitive physics benchmarks, we show that visual mental simulation can improve prediction over observed-only and language-based reasoning, and analyze how its benefit depends on the VLM and the quality of the generated future.

2 Related Work

Vision-language models. Modern vision-language models (VLMs) combine strong visual perception with language-based reasoning, enabling broad multimodal capabilities such as spatial reasoning, fine-grained perception, and video understanding DeepMind (2025); Gemini Team (2025); OpenAI (2025); Bai et al. (2025); Wang et al. (2025); Tong et al. (2024); Li et al. (2023); Liu et al. (2023); Yang et al. (2026). Inference-time methods such as chain-of-thought prompting and test-time compute scaling further improve reasoning by allocating additional computation to intermediate reasoning trajectories Muennighoff et al. (2025); Wei et al. (2022); Wang et al. (2022); Zhang et al. (2023). However, such scaling primarily expands in language space and remains grounded in the observed input. In contrast, our work explores allocating test-time computation in visual-dynamic space by generating candidate future states of the scene before answering.

Video models as world models and simulators. Building on the success of large-scale video generation models Brooks et al. (2024); Wan et al. (2025); Blattmann et al. (2023); Yang et al. (2024), recent work has explored using video models as world models that can simulate future states of the environment conditioned on past observations Valevski et al. (2025); Bruce et al. (2024); Chen et al. (2024); Song et al. (2025); Chen et al. (2025). Such rollouts have supported planning and control in robotics and decision making Qi et al. (2025); Du et al. (2023; 2024). Rather than

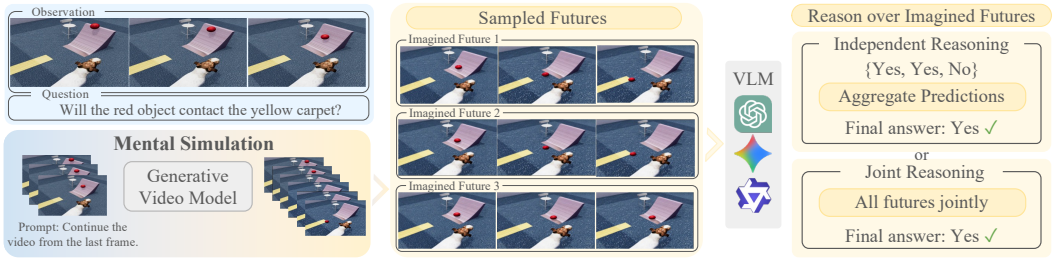


Figure 2: **Method overview.** Given an observed video segment and a question, we generate multiple plausible future rollouts using a generative video model conditioned on the observation. These sampled futures are then provided to a vision-language model (VLM) as additional visual evidence to answer the question. The VLM can reason over each rollout independently and aggregate the resulting predictions, or jointly reason over multiple imagined futures to produce the final answer.

treating these rollouts as literal predictions of the future, we view them as stochastic hypotheses that serve as intermediate visual evidence for downstream reasoning.

Generative models for visual reasoning. A growing line of work uses generative models to support reasoning by producing intermediate visual states at test time. Image-based approaches generate diagrams, sketches, or intermediate states to aid multimodal reasoning Chern et al. (2025); Su et al. (2025); Li et al. (2025); Liu et al. (2022). Recent work also studies video generation as a reasoning medium Tong et al. (2026) and controllable world models that synthesize views for spatial reasoning from a single image Yang et al. (2025). We focus on future-dependent physical events beyond an observed video, using generated rollouts as explicit evidence for prediction.

Physical scene understanding and intuitive physics. Physical scene understanding concerns inferring the physical state of a scene and predicting how it will evolve under forces, collisions, and contact Battaglia et al. (2013); Wu et al. (2015); Watters et al. (2017); Agrawal et al. (2016). A common formulation is intuitive physics, where models must forecast future outcomes from partial observations Yi et al. (2020); Motamed et al. (2025); Bear et al. (2021); Bordes et al. (2025); Riochet et al. (2018). We instantiate our framework in this clean yet challenging setting and evaluate on CLEVRER Yi et al. (2020), Physion Bear et al. (2021), and Physics-IQ (repurposed) Motamed et al. (2025), spanning synthetic and real-world interactions and directly testing questions about future events from partial observations. By contrast, IntPhys and IntPhys 2 use a violation-of-expectation formulation that asks whether complete videos are physically possible Bordes et al. (2025); Riochet et al. (2018), which is less directly aligned with our goal of forecasting explicit future events.

3 Method

Vision-language models (VLMs) can reason over observed visual evidence, but future-dependent questions require extrapolating beyond what is directly visible. We introduce a test-time simulate-then-reason framework that makes this extrapolation explicit in visual space: rather than relying only on the VLM to infer future dynamics internally, we use a generative video model to produce multiple plausible future continuations. These imagined continuations are then provided to the VLM as additional visual evidence for answering the question (Figure 2).

3.1 Problem Formulation

Let $V_{0:T}$ denote an observed video segment consisting of frames from time 0 to T. Given a question q (e.g., "What will happen next?") and answer options $\mathcal{Y} = \{y_1, \ldots, y_m\}$, the goal is to predict the correct answer $\hat{y} \in \mathcal{Y}$. In some scenarios, the correct answer depends on events occurring after time T and is therefore not directly observable in the input segment. Solving the task requires anticipating how the scene may evolve after time T (e.g., object motion and interactions).

Observed-only baseline. A standard VLM baseline predicts the answer directly from the observed segment:

$$\hat{y}_{\text{base}} = f_{\text{VLM}}(V_{0:T}, q), \qquad \hat{y}_{\text{base}} \in \mathcal{Y},$$

where $f_{\text{VLM}}(\cdot)$ denotes the answer produced by the VLM and parsed as an option in $\mathcal{Y}$; invalid answers count as incorrect. In this formulation, the VLM must implicitly infer how the scene evolves

beyond the observed frames in order to determine which answer is consistent with the unobserved future.

Language-space reasoning baseline. We also consider two VLM-only language-space reasoning baselines, both of which use only the observed video as visual input and prompt models to reason internally before answering. The first uses a structured prompt that organizes the observed scene in terms of relevant objects, motions, and interactions before answering. The second asks the model to internally simulate how these dynamics may evolve next. We denote such prompts by c_{lang} and write $\hat{y}_{\text{lang}} = f_{\text{VLM}}(V_{0:T}, q; c_{\text{lang}})$. These baselines use only the observed video as visual input and request the final answer without a printed explanation. In contrast, the written-future setting explicitly asks the model to describe the predicted future before answering.

3.2 Mental Simulation via Video Rollout

Humans often reason about outcomes by mentally simulating possible futures before making a decision. Inspired by this idea, we externalize such future simulation for VLMs by introducing an explicit generative step that produces candidate continuations of the scene.

Video rollout generation. We use a diffusion-based generative video model G to sample imagined future rollouts conditioned on the observed segment:

$$\hat{V}_{T+1:T+\Delta} \sim p_G(\cdot \mid V_{0:T}, c),$$

where c is a textual prompt instructing the model to continue the video (e.g., "continue the video from the last frame"), and Δ denotes the rollout horizon. Because future evolution may be uncertain and generation is stochastic, we draw K continuations $\hat{V}^{(k)}_{T+1:T+\Delta}$ using different random seeds. Each rollout represents a candidate continuation of the observed scene.

Conditioning on observed dynamics. We condition video generation on the full observed segment rather than only the current frame. Unlike image-to-video generation, video continuation models Chen et al. (2025) can use multiple observed frames as temporal context. This is particularly important in dynamic scenes, where motion cues (e.g., direction and speed) and upcoming interactions are often difficult to infer from a single frame or from text prompts alone. During generation, the observed history guides the prediction of future frames, encouraging the continuation to remain consistent with the preceding dynamics. This helps maintain scene layout, object identities, and motion cues, producing temporally coherent continuations of the scene.

The imagined rollouts are treated as plausible future hypotheses rather than ground-truth predictions. We then provide these rollouts to the VLM as additional visual evidence for its final prediction.

3.3 Future-Aware Reasoning with Imagined Rollouts

Imagined future rollouts can provide useful visual evidence, but an individual rollout may contain artifacts that are noisy or misleading. We therefore reason over multiple generated continuations rather than relying on a single imagined future.

For each sampled rollout k, we concatenate the observed segment with the generated continuation to form an extended video $\hat{V}^{(k)} = [V_{0:T}, \hat{V}^{(k)}_{T+1:T+\Delta}]$. Given multiple imagined futures, the VLM can integrate them in two natural ways: by reasoning over each rollout independently and aggregating the resulting predictions, or by reasoning jointly over all rollouts in a single inference.

Independent reasoning. We run the VLM independently on each $\hat{V}^{(k)}$ together with the question q, and obtain a rollout-level predicted answer $\hat{y}^{(k)} = f_{\text{VLM}}(\hat{V}^{(k)}, q)$. We aggregate predictions across rollouts using majority voting $\hat{y}_{\text{vote}} = \text{mode}\left(\hat{y}^{(1)}, \ldots, \hat{y}^{(K)}\right)$. This allows each rollout to contribute one hypothesis about the future while reducing reliance on any single generated continuation.

Joint reasoning. Alternatively, the VLM receives the observed prefix alongside all rollout-extended videos and produces a single prediction $\hat{y}_{\text{joint}} = f_{\text{VLM}}(V_{0:T}, \hat{V}^{(1)}, \ldots, \hat{V}^{(K)}, q; c_{\text{joint}})$, where c_{joint} prompts the model to compare the imagined futures before answering. This allows the model to directly integrate visual evidence across multiple possible scene evolutions.

In both cases, generated rollouts serve as visual hypotheses of the future. Independent reasoning aggregates separate rollout answers, whereas Joint reasoning considers the rollouts together with

Algorithm 1 Mental Simulation for Future-Aware VLM Reasoning

Require: Observed video $V_{0:T}$, question q, options $\mathcal{Y}$, generator G, VLM f_{VLM}, mode r, prompts c, c_{joint}, horizon Δ, and rollout count K.
Ensure: Predicted answer $\hat{y} \in \mathcal{Y}$.

(1) Mental simulation via video rollout:
for $k = 1$ to K **do**
  Sample $\hat{V}^{(k)}_{T+1:T+\Delta} \sim p_G(\cdot \mid V_{0:T}, c)$ // different seed per k
  Form extended video $\hat{V}^{(k)} \leftarrow [V_{0:T}, \hat{V}^{(k)}_{T+1:T+\Delta}]$
end for
(2) Future-aware reasoning:
if $r =$ INDEPENDENT **then**
  $\hat{y}^{(k)} \leftarrow f_{\text{VLM}}(\hat{V}^{(k)}, q)$ for each k.
  $\hat{y} \leftarrow \text{mode}\left(\hat{y}^{(1)}, \ldots, \hat{y}^{(K)}\right)$ // majority vote
else if $r =$ JOINT **then**
  $\hat{y} \leftarrow f_{\text{VLM}}\left(V_{0:T}, \hat{V}^{(1)}, \ldots, \hat{V}^{(K)}, q; c_{\text{joint}}\right)$ // joint reasoning
end if
return $\hat{y}$

observed context. Algorithm 1 summarizes the pipeline, which allocates test-time computation to generating and evaluating candidate visual futures.

4 EXPERIMENTS

We evaluate whether augmenting VLMs with mental simulation improves their ability to reason about future physical outcomes on three intuitive physics benchmarks. We compare observed-only inference, language-space reasoning, and visual simulation with imagined future rollouts, and analyze how performance depends on rollout integration, video generator, and rollout quality.

4.1 BENCHMARKS

We evaluate on three benchmarks.

Physion Bear et al. (2021). A simulated benchmark of physical interactions. We evaluate on 1,035 questions spanning 7 scenario types (dominoes, roll, collide, drop, link, support, contain). Each example consists of an observed video segment and a binary question about whether a future contact event will occur.

CLEVRER Yi et al. (2020). A simulated benchmark of multi-object interactions. We evaluate on 1,000 predictive questions from the validation split, where the model observes a 5-second video and predicts what happens in the following 2 seconds (e.g., which objects will collide next).

Physics-IQ (repurposed) Motamed et al. (2025). A dataset of real-world physical interaction videos that we repurpose into a predictive multiple-choice QA format. We generate question–answer options from video descriptions using GPT, and a human annotator verifies the correct answer. Our evaluation uses 66 human-reviewed predictive questions, one per scene.

4.2 MODELS AND SETTINGS

VLM backbones. We evaluate a broad set of open-source and proprietary VLMs selected to cover multiple model families, capability levels, and inference configurations. Our main comparison includes Qwen3-VL Bai et al. (2025), Kimi-K3 Team et al. (2026), Gemini Gemini Team (2025), and GPT OpenAI (2026) families. We additionally evaluate GPT-6 Astra on targeted subsets due to its higher inference cost. Full model and inference details are provided in the supplement.

Inference settings. To understand how different forms of test-time reasoning affect future-dependent prediction, we compare inference from the observed video alone, explicit reasoning in language, and reasoning augmented with generated future visual evidence.

Table 1: **Observed, language, and visual reasoning across VLMs.** Accuracy (%) on Physion (1,035 questions), CLEVRER (1,000), and Physics-IQ (66). Avg. weights the three datasets equally; bold/underline mark the best/second-best values per model and block, including ties. Visual simulation can improve prediction, with gains varying across models and datasets.

Space	Setting	Qwen3-VL			Kimi-K3			Gemini			GPT	
		8B	32B	235B	low	gw	max	2.5-Pro	3-Fl	3.7-Fl	5.2	5.6
Avg.												
Observed	Single	50.0	56.6	55.9	66.5	67.4	74.3	63.3	66.1	75.5	67.5	70.1
	vote@3	50.6	56.5	55.1	69.0	70.4	75.0	65.5	66.2	77.1	68.7	70.1
Language	Structured	50.7	56.5	55.9	69.0	67.9	75.2	62.0	66.5	76.8	68.6	71.7
	Structured vote@3	50.6	57.0	54.8	70.1	69.9	74.7	64.9	66.4	76.3	68.5	70.7
	Future	49.4	58.2	57.4	68.5	68.8	75.3	60.9	65.1	76.6	66.3	69.3
	Future vote@3	49.5	58.0	56.4	68.9	71.2	75.0	63.1	65.3	78.2	67.1	68.8
Visual	Independent	**62.2**	**69.5**	**70.2**	**76.3**	74.8	76.4	**74.0**	74.7	73.1	72.0	75.2
	Joint	50.6	65.8	65.7	76.2	**77.5**	**79.2**	73.2	**77.0**	**82.5**	**75.1**	**76.7**
Physion ($n = 1{,}035$)												
Observed	Single	47.4	52.6	51.1	56.7	56.7	67.2	55.7	55.9	66.1	60.3	62.4
	vote@3	47.4	52.6	51.2	58.1	58.9	66.7	58.1	55.9	66.2	60.9	62.7
Language	Structured	49.2	50.8	50.6	57.3	56.1	64.6	56.4	55.5	64.7	61.8	62.5
	Structured vote@3	49.1	51.0	50.6	57.8	58.7	64.5	56.5	55.3	66.5	62.6	**64.5**
	Future	48.5	54.0	52.5	58.3	58.1	**67.5**	56.3	56.1	65.4	62.1	60.4
	Future vote@3	48.5	54.0	52.6	58.1	61.8	**67.5**	56.7	55.4	**67.0**	61.5	61.4
Visual	Independent	**56.4**	**60.9**	60.4	59.3	59.9	61.6	**60.1**	56.8	58.6	60.5	60.9
	Joint	48.9	60.3	**62.6**	**62.6**	**64.4**	63.7	59.7	**64.4**	64.0	**63.0**	61.6
CLEVRER ($n = 1{,}000$)												
Observed	Single	63.2	65.8	68.1	86.7	86.4	89.1	72.2	66.7	89.1	80.1	85.8
	vote@3	63.4	65.5	68.6	86.7	87.0	90.1	74.7	67.0	89.3	81.5	85.5
Language	Structured	63.4	67.2	68.5	86.0	86.9	88.3	71.9	66.9	88.4	80.2	84.3
	Structured vote@3	63.2	67.1	68.4	85.9	87.3	89.8	74.5	68.2	89.8	82.3	85.3
	Future	63.3	65.9	68.2	86.6	86.3	88.7	73.4	66.3	90.1	80.8	85.4
	Future vote@3	63.5	66.9	68.3	86.4	86.5	89.3	74.9	67.7	90.4	80.7	85.9
Visual	Independent	**74.2**	**85.5**	**88.2**	90.9	93.3	94.8	**89.1**	**92.9**	92.7	90.4	93.4
	Joint	57.5	73.4	78.3	**94.8**	**93.7**	**96.7**	85.6	89.4	**97.2**	**94.2**	**94.2**
Physics-IQ ($n = 66$)												
Observed	Single	39.4	51.5	48.5	56.1	59.1	66.7	62.1	75.8	71.2	62.1	62.1
	vote@3	40.9	51.5	45.5	62.1	65.2	68.2	63.6	75.8	75.8	63.6	62.1
Language	Structured	39.4	51.5	48.5	63.6	60.6	72.7	57.6	**77.3**	77.3	63.6	68.2
	Structured vote@3	39.4	53.0	45.5	66.7	63.6	69.7	63.6	75.8	72.7	60.6	62.1
	Future	36.4	54.5	51.5	60.6	62.1	69.7	53.0	72.7	74.2	56.1	62.1
	Future vote@3	36.4	53.0	48.5	62.1	65.2	68.2	57.6	72.7	77.3	59.1	59.1
Visual	Independent	**56.1**	62.1	**62.1**	**78.8**	71.2	72.7	72.7	74.2	68.2	65.2	71.2
	Joint	45.5	**63.6**	56.1	71.2	**74.2**	**77.3**	**74.2**	**77.3**	**86.4**	**68.2**	**74.2**

- **Observed-only inference.** *Observed* predicts the answer directly from the observed segment $V_{0:T}$. As a repeated-inference control, *Observed vote@3* runs the same inference three times and aggregates the resulting predictions by majority vote.
- **Language-space reasoning.** The VLM still receives only the observed video; however, it is now prompted to reason about the scene before answering. *Language–Structured* organizes the observed scene in terms of relevant objects, motions, and interactions. *Language–Future* asks the model to simulate how these dynamics may evolve. We evaluate both settings with a single inference and with *vote@3*.
- **Visual mental simulation.** We generate $K = 3$ future rollouts as additional visual evidence. *Visual–Independent* evaluates each rollout separately and majority-votes the predictions, whereas *Visual–Joint* jointly reasons over all three rollouts plus a separately observed video. Each rollout contains context frames followed by generated future frames.

Video generation models. Our main experiments use Large Video Planner (LVP) Chen et al. (2025) to generate imagined future rollouts for all VLM backbones. To study how downstream reasoning

Table 2: **Future reasoning in different spaces.** Accuracy (%) on 100 questions each from Physion and CLEVRER. Visual simulation achieves the highest accuracy for every model on both datasets.

		Physion ($n = 100$)				**CLEVRER** ($n = 100$)			
Space	Setting	**Qwen235B**	**Gemini3.7**	**GPT5.6**	**GPT6**	**Qwen235B**	**Gemini3.7**	**GPT5.6**	**GPT6**
Observed	Single	55.0	61.0	52.0	58.0	67.0	91.0	81.0	90.0
	Joint control	45.0	62.0	57.0	63.0	68.0	88.0	81.0	90.0
Language	Structured	56.0	56.0	54.0	58.0	69.0	90.0	80.0	91.0
	Future	58.0	62.0	47.0	61.0	71.0	89.0	81.0	92.0
	Future (written)	61.0	50.0	54.0	57.0	72.0	87.0	78.0	90.0
Tracking	Full tracks	45.0	48.0	51.0	65.0	61.0	89.0	80.0	92.0
	Moving tracks	54.0	47.0	53.0	60.0	61.0	84.0	79.0	90.0
	Full summary	55.0	45.0	49.0	66.0	64.0	87.0	77.0	87.0
	Moving summary	52.0	48.0	48.0	64.0	60.0	84.0	80.0	91.0
Visual	Independent	**71.0**	71.0	**72.0**	72.0	**87.0**	91.0	91.0	92.0
	Joint	66.0	**74.0**	68.0	**80.0**	82.0	**96.0**	**95.0**	**96.0**

depends on the video generator, we additionally compare LVP with Cosmos3-Super Agarwal et al. (2026) and LongCat Team et al. (2025). We adapt each generator's supported continuation interface while keeping the evaluation questions and VLM answering procedure fixed. Additional details are provided in the supplementary material.

4.3 Quantitative Results

Table 1 compares observed-only inference, language-space reasoning, and visual mental simulation across VLM families and benchmarks.

Visual simulation provides information beyond language-space reasoning. Self-consistency and language-space reasoning improve performance in some settings, showing that stronger inference alone can recover part of the gap. However, generated future rollouts often provide substantially larger gains. On CLEVRER, for example, Qwen3-VL-32B improves from 65.8% with *Observed* to 85.5% with *Visual–Independent*, while GPT-5.6 improves from 85.8% to 94.2% with *Visual–Joint*. Similar gains appear for several models on Physion. Table 2 further compares visual simulation with explicit language-based future reasoning and tracking information on subsets. These results suggest that explicitly constructing possible future visual states can provide useful evidence that is not always recovered by reasoning over the observed video alone.

VLM capability affects how imagined futures are integrated. More capable VLMs often achieve higher observed-only accuracy, yet explicit visual simulation can still provide substantial gains for capable models. More interestingly, stronger models are often better able to reason jointly over multiple imagined futures, while smaller models can benefit more from considering each rollout independently. For example, Qwen3-VL-8B reaches 74.2% on CLEVRER with *Visual–Independent* but 57.5% with *Visual–Joint*, whereas GPT-5.6 and Gemini-3.7-Flash achieve their best CLEVRER results with joint reasoning.

Gains are most consistent on CLEVRER, vary across VLMs and integration strategies on Physion, and provide exploratory real-world evidence on the smaller Physics-IQ evaluation. We next vary the generator, compare generated and ground-truth futures, and examine the association between human-rated rollout quality and answer accuracy.

4.4 Analysis and Ablations

To keep additional inference and human-annotation costs manageable, we randomly select 100 questions from each of Physion and CLEVRER and reuse these subsets for the generator, prompting, and rollout-quality analyses.

Visual simulation with alternative generators. We evaluate whether useful future evidence can also be obtained from generators beyond LVP. We adapt each generator's supported continuation interface to the same observed scenes and evaluate three continuations using the same joint VLM answering procedure in Table 3. The table shows both observed controls and reports gains over the higher of their two accuracies. Cosmos3-Super and LongCat also improve accuracy in several

Table 3: **Visual simulation with different future generators.** Accuracy (%) on Physion and CLEVRER (100 questions each) and Physics-IQ (66). Parentheses show gains over the higher accuracy of Observed Single and Joint control in each column. Bold marks the best non-GT result, including ties. Alternative generators can also provide useful future evidence, with benefits varying across models and datasets.

		Physion ($n = 100$)				**CLEVRER** ($n = 100$)				**Physics-IQ** ($n = 66$)			
Evidence	Setting	**Qwen 235B**	**Gemini 3.7**	**GPT 5.6**	**GPT 6**	**Qwen 235B**	**Gemini 3.7**	**GPT 5.6**	**GPT 6**	**Qwen 235B**	**Gemini 3.7**	**GPT 5.6**	**GPT 6**
Observed	Single	55.0	61.0	52.0	58.0	67.0	91.0	81.0	90.0	47.0	78.8	68.2	78.8
	Joint control	45.0	62.0	57.0	63.0	68.0	88.0	81.0	90.0	48.5	69.7	65.2	74.2
Generated future	LVP	**66.0** (+11.0)	**74.0** (+12.0)	**68.0** (+11.0)	**80.0** (+17.0)	**82.0** (+14.0)	**96.0** (+5.0)	**95.0** (+14.0)	**96.0** (+6.0)	**60.6** (+12.1)	**81.8** (+3.0)	**75.8** (+7.6)	**80.3** (+1.5)
	Cosmos	53.0 (−2.0)	68.0 (+6.0)	60.0 (+3.0)	63.0 (0.0)	73.0 (+5.0)	94.0 (+3.0)	86.0 (+5.0)	95.0 (+5.0)	45.5 (−3.0)	62.1 (−16.7)	60.6 (−7.6)	66.7 (−12.1)
	LongCat	52.0 (−3.0)	59.0 (−3.0)	59.0 (+2.0)	63.0 (0.0)	71.0 (+3.0)	92.0 (+1.0)	89.0 (+8.0)	90.0 (0.0)	42.4 (−6.1)	78.8 (0.0)	71.2 (+3.0)	77.3 (−1.5)
GT future		80.0 (+25.0)	97.0 (+35.0)	92.0 (+35.0)	99.0 (+36.0)	—	—	—	—	74.2 (+25.8)	87.9 (+9.1)	78.8 (+10.6)	80.3 (+1.5)

Table 4: **Human-rated rollout quality on Physion.** Single-rollout accuracy (%) across VLMs on the same 46 questions with both video ratings. Gap is Plausible − Implausible in percentage points. Physically plausible rollouts have higher point accuracy.

	Qwen3-VL			**Kimi-K3**			**Gemini**			**GPT**		
Rating	8B	32B	235B	low	gw	max	2.5-Pro	3-Fl	3.7-Fl	5.2	5.6	6
Physion ($n = 46$)												
Implausible	60.9	58.7	71.7	80.4	82.6	78.3	69.6	78.3	82.6	65.2	78.3	80.4
Plausible	73.9	73.9	80.4	82.6	84.8	87.0	80.4	82.6	84.8	80.4	82.6	84.8
Gap (pp)	+13.0	+15.2	+8.7	+2.2	+2.2	+8.7	+10.9	+4.3	+2.2	+15.2	+4.3	+4.3

model–dataset settings, although the gains vary. LVP gives the highest accuracy across Physion and CLEVRER for the four evaluated VLMs, while Physics-IQ is more mixed. These results show that the benefit can extend beyond a single generator, but useful future evidence is not guaranteed by adding generated video. Differences may reflect both the generated continuations and how the VLM interprets them.

Strong VLMs still benefit from informative future evidence. We evaluate GPT-6 Astra with ground-truth future video to assess its ability to use accurate future evidence. On Physion, GPT-6 improves from 63.0% with the observed joint control to 80.0% with LVP, and reaches 99.0% with the ground-truth future. On CLEVRER, LVP improves accuracy from 90.0% to 96.0%. On Physics-IQ, both LVP and ground-truth futures reach 80.3%, compared with 78.8% for the stronger observed control. Thus, weaker gains from current generators do not imply that future visual evidence is no longer useful for stronger VLMs; substantial headroom remains on Physion when the future itself is more accurate.

Rollout quality vs accuracy. One human annotator assessed the physical plausibility of generated LVP rollouts for the 100-question Physion subset without viewing the VLM answers. To avoid comparing different mixtures of easier and harder questions, we retain the 46 questions with both a physically plausible and a physically implausible rollout. We select one video from each rating group using a fixed rule independent of VLM answers and reuse the same pairs across all twelve configurations, with equal question weights. Table 4 reports single-rollout QA accuracy on Physion: physically plausible rollouts have higher point accuracy in all twelve configurations, with gaps of 2.2–15.2 percentage points. Additional rollout-quality results are provided in Sec. B.1.

4.5 Qualitative Results

Figure 3 shows examples where the VLM answers incorrectly from the observed segment alone, but correctly when given an imagined future rollout. In the Physics-IQ example, the rollout shows the white block remaining on the rotating platform; in the CLEVRER example, it shows the object interaction that occurs after the observed segment. The CLEVRER example further illustrates how the added future evidence changes the VLM's answer: from the observed clip, the model predicts an earlier cylinder–sphere interaction, while the imagined continuation shows the cube–cube collision, leading to the correct answer. Overall, these examples show how video generation can extend a partial observation into an explicit candidate future that the VLM can directly reason over. Additional

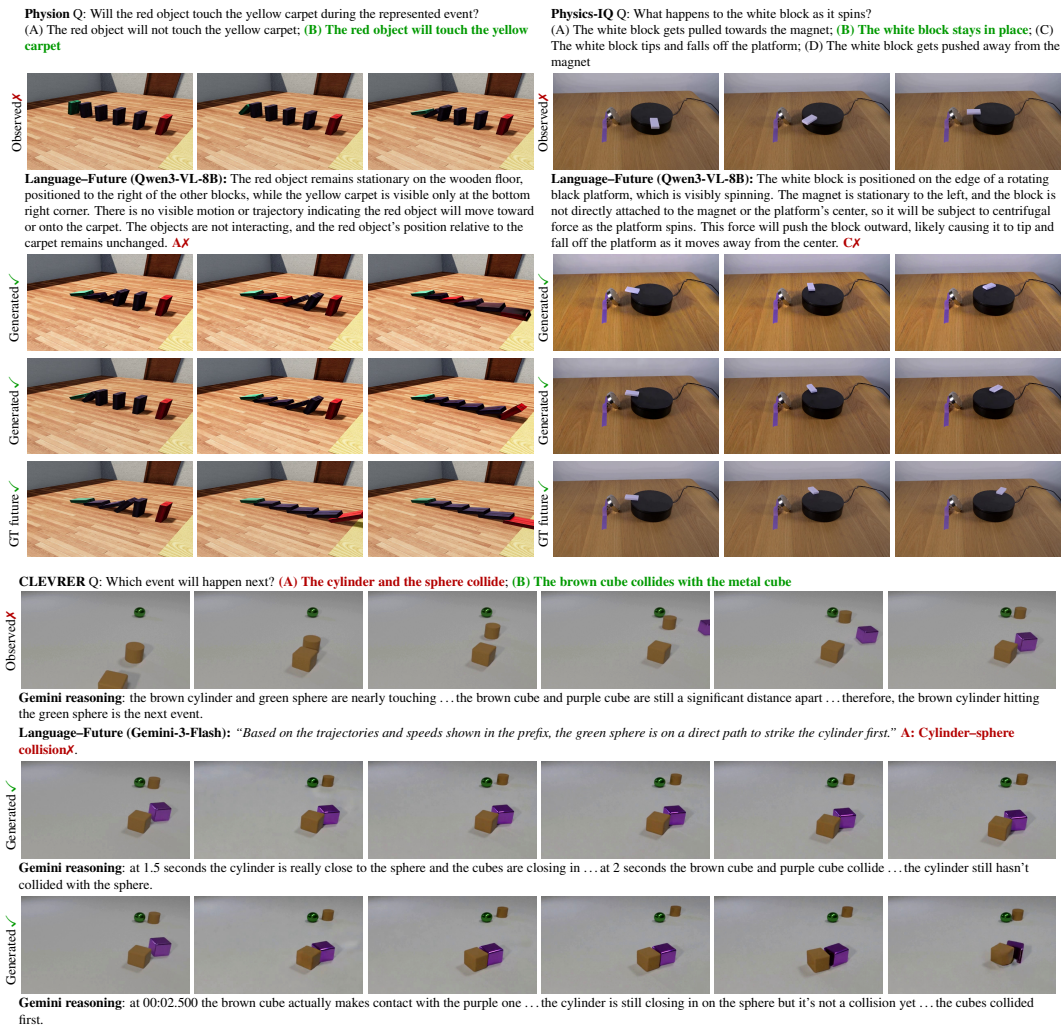


Figure 3: **Qualitative examples from Physion (top left), Physics-IQ (top right), and CLEVRER (bottom).** Each shows the observed segment and two generated futures; GT futures are not available for CLEVRER. Observed-only and single-rollout Qwen3-VL-8B predictions are marked with ✓/✗. For CLEVRER, shortened Gemini-3-Flash traces show how generated future evidence changes the predicted next event from a cylinder–sphere to a cube–cube collision; its answers agree with Qwen3-VL-8B. Language–Future rows show 2-3 sentence forecasts from the observed video.

qualitative examples, failure cases, and rollout visualizations are provided in the supplementary material and project webpage.

5 Conclusion

We introduce a simulate-then-reason framework that augments VLMs with generated future videos at test time. Across Physion, CLEVRER, and the repurposed Physics-IQ evaluation set, imagined rollouts can improve future-dependent prediction over baselines that rely only on observed videos or use language-based reasoning. Our analyses further show that these gains depend both on rollout quality and how the VLM uses the generated visual context. On Physion, the remaining gap to ground-truth futures suggests room for improvement in both. Together, these findings motivate explicit visual simulation as a complementary direction for test-time reasoning about events beyond the observed input.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, provide feedback on research methodology and experiments, implement methods, and clean and reformat datasets. We have not used generative AI tools to assist with translation, propose or refine hypotheses, help develop theoretical models or conceptual frameworks, support qualitative or thematic data analysis, or interpret results. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures, create or edit code, summarize or analyze existing literature, and edit the research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was human-verified and tested for correctness by the authors, and the final passes of the paper were completed by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800. 2, 7

Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. *Advances in neural information processing systems*, 29, 2016. 3

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1, 2, 5

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. doi: 10.1073/pnas.1306572110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1306572110. 1, 3

Daniel Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Fan-Yun Sun, Fei-Fei Li, Nancy Kanwisher, Josh Tenenbaum, Dan Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d09bf41544a3365a46c9077ebb5e35c3-Abstract-round1.html. 2, 3, 5

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. URL https://arxiv.org/abs/2311.15127. 2

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849. 3

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators. 2

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott E. Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html. 2

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024. 2

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner enables generalizable robot control, 2025. URL https://arxiv.org/abs/2512.15840. 2, 4, 6

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images, 2025. URL https://arxiv.org/abs/2505.22525. 3

Google DeepMind. Gemini 3, 2025. https://deepmind.google/models/gemini/ [Accessed: (March 2026)]. 1, 2

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html. 2

Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Pack Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=9pKtcJcMP3. 2

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261. 2, 5

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 3

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023. 2

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*, 2022. 3

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. 2, 3, 5

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1025. URL https://doi.org/10.18653/v1/2025.emnlp-main.1025. 1, 2

OpenAI. Introducing gpt-5.2, 2025. https://openai.com/index/introducing-gpt-5-2/ [Accessed: (March 2026)]. 1, 2

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026. https://openai.com/index/gpt-5-6/ [Accessed: (September 21, 2026)]. 2, 5

Han Qi, Haocheng Yin, Aris Zhu, Yilun Du, and Heng Yang. Strengthening generative robot policies through predictive world modeling, 2025. URL https://arxiv.org/abs/2502.00622. 2

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018. 3

Daniel L Schacter, Donna Rose Addis, and Randy L Buckner. Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 8:657–661, 2007. doi: 10.1038/nrn2213. 1

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/song25b.html. 2

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers, 2025. URL https://arxiv.org/abs/2506.23918. 3

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. 5

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report, 2025. URL https://arxiv.org/abs/2510.22200. 7

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 03-07, 2026*, 2026. 3

Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Vi8AepAXGy. 2

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=P8pqeEkn1H. 2

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314. 2

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2

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. *arXiv preprint arXiv:2203.11171*, 2022. 2

Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. *Advances in neural information processing systems*, 30, 2017. 3

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html. 1, 2

Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in neural information processing systems*, 28, 2015. 3

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*, 2026. URL https://openreview.net/forum?id=rBFDvZu6pb. 2

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=L2W4wQsNkY. 3

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HkxYzANYDB. 2, 3, 5

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 2

A Implementation Details

A.1 Evaluation sets and inference settings

The broad comparison uses 1,035 Physion questions, 1,000 CLEVRER questions, and 66 human-reviewed Physics-IQ questions, one per scene. We randomly select 100 questions from each of Physion and CLEVRER to keep additional inference and human annotation within budget, and reuse these subsets across the generator, prompting, and rollout-quality analyses.

Observed Single predicts from the observed video alone. The Structured and Future prompts ask the VLM to identify physical relations or anticipate future motion, respectively, and request only a final answer. Their vote@3 versions aggregate three answers. The written-future setting additionally requests a forecast before the answer. Full tracks provide point trajectories extracted from the observed clip; Moving tracks retain only trajectories passing the motion-selection criteria. Full and Moving summaries express the corresponding trajectory information as text.

Visual Independent evaluates each of three generated continuations separately and votes over the answers; each call receives the candidate clip without a separately attached observed prefix. Visual Joint receives the observed prefix and all three continuations together. Its observed joint control uses the corresponding joint-answer procedure with no generated continuations. Invalid answers count as incorrect, and voting uses a fixed tie-breaking rule. Shared generator/reasoning table entries reuse identical predictions. Written-future and tracking settings marked with † come from additional runs.

A.2 Video continuation and generator evaluation

We adapt LVP, Cosmos3-Super, and LongCat to continue the same observed scenes. Each supplies three continuations, which are evaluated on the same questions with the same joint VLM answering procedure. We use model-specific conditioning interfaces and continuation prompts; the comparison concerns these practical generation configurations rather than isolating architecture under identical generation budgets.

LVP produces 49-frame clips at 832×480 and 16 FPS. Physion and Physics-IQ use 17 context frames followed by 32 generated frames; CLEVRER uses 25 context frames followed by 24 generated frames. Its continuation instructions encourage a fixed camera, preserved objects, and physically plausible motion.

Cosmos3-Super uses five conditioning frames and a shared, scene-independent prompt asking it to continue the scene while preserving objects and camera position. The evaluated continuation contains 22 future frames at 8 FPS and 832×480. Cosmos uses 35 steps, guidance 6, and flow shift 10. LongCat uses a scene-dependent continuation prompt describing the observed scene and requesting consistent objects, camera, and motion; the prompt is held fixed across its three seeds. Continuation configurations were checked on pilot examples before the full generator evaluation, and all fixed-seed outputs are retained.

Generated-video prompts instruct the VLM to answer from the represented continuation, allowing for artifacts and without predicting beyond its end. The broad LVP and human-quality runs use LVP-specific wording; the generator comparison uses generator-neutral wording. Questions and answer options are unchanged.

A.3 Human rollout-quality assessment

One human annotator assessed the physical plausibility of each rollout without viewing the VLM answers. A rollout is rated physically plausibility when its motion and interactions are physically plausible; it need not reproduce the actual future outcome. Physion and CLEVRER each have three annotated rollouts for 100 contexts; Physics-IQ has annotated rollouts for 30 of the evaluated questions. The analysis retains questions with both a physically plausibility and a physically implausibility rollout: 46 Physion, 52 CLEVRER, and 10 Physics-IQ questions.

A fixed hash selects one video from each rating group independently of VLM answers, and all twelve configurations use the same video pairs with equal question weights. Matching controls question composition; gaps describe an association, not an isolated effect of physical plausibility.

B Additional Results

The supplementary results provide additional rollout-quality analyses and examine observed controls and uncertainty in the main comparisons.

B.1 Human-rated rollout quality: additional details and datasets

The Physion analysis in Table 4 and the CLEVRER analysis below use the same starting 100-question subsets as the generator and prompting analyses. Their single-rollout prompts retain the earlier wording "one fallible LVP candidate future," whereas the generator comparison uses "one generated video" and generator-neutral evidence labels. The shared single-rollout instruction otherwise asks which answer the continuation supports, allows for generation artifacts, and forbids predicting beyond its end; human ratings are not provided to the VLM.

On CLEVRER, 52 of 100 annotated questions contain both rollout ratings. Physically plausible rollouts have higher point accuracy in seven of twelve configurations, with four ties and one decrease. We retain this more mixed pattern alongside the Physion results for completeness.

The selected 66-question Physics-IQ evaluation contains 30 questions with human-rated rollouts. Only 10 questions, each from a different scene, have both a physically plausible and a physically implausible rollout. We use the same twelve configurations, LVP-specific prompts, and fixed, answer-independent pair selection as Table 4. Matching keeps question identity and weight equal between rating groups, reducing confounding from different mixtures of easier and harder questions. It does not isolate physical plausibility from other differences between videos, and the findings apply to the mixed-quality subset.

Table 5: **Human-rated rollout quality on CLEVRER and Physics-IQ.** Single-rollout accuracy (%) on 52 CLEVRER and 10 Physics-IQ questions with both video ratings, using the same pairs across twelve configurations. Gap is Plausible − Implausible in percentage points; colors indicate direction, not significance. The pattern is more mixed than on Physion.

	Qwen3-VL			**Kimi-K3**			**Gemini**			**GPT**		
Rating	8B	32B	235B	low	gw	max	2.5-Pro	3-Fl	3.7-Fl	5.2	5.6	6
CLEVRER ($n = 52$)												
Implausible	78.8	84.6	84.6	84.6	84.6	86.5	76.9	90.4	90.4	82.7	90.4	90.4
Plausible	76.9	84.6	90.4	86.5	94.2	96.2	82.7	90.4	90.4	82.7	92.3	92.3
Gap (pp)	−1.9	0.0	+5.8	+1.9	+9.6	+9.6	+5.8	0.0	0.0	0.0	+1.9	+1.9
Physics-IQ ($n = 10$)												
Implausible	70.0	90.0	90.0	90.0	80.0	80.0	80.0	80.0	70.0	70.0	70.0	70.0
Plausible	70.0	100.0	80.0	90.0	90.0	90.0	100.0	90.0	90.0	70.0	100.0	80.0
Gap (pp)	0.0	+10.0	−10.0	0.0	+10.0	+10.0	+20.0	+10.0	+20.0	0.0	+30.0	+10.0

B.2 Observed controls and uncertainty in the full evaluation

Table 6 extends the main comparison with Observed Joint control, confidence intervals, and two visual-gain rows beneath each dataset block. It retains the same eleven configurations and evaluation questions, including the corrected Gemini-3-Flash and GPT-5.2 predictions. Δ Independent subtracts Observed vote@3 from Visual Independent; Δ Joint subtracts Observed Joint control from Visual Joint.

We compute pointwise 95% percentile intervals from 10,000 bootstrap resamples of scenes, sharing each resample across models and settings. The evaluated populations have one question per scene. Paired gains use the same resampled scenes on both sides; Avg. weights the three datasets equally, with independent resampling across datasets. Intervals condition on the saved predictions and selected questions, do not measure inference-rerun variability, and are not adjusted for multiple comparisons. Gains and interval endpoints are computed before rounding. Kimi low/gateway/max denote inference configurations; GPT-5.6 denotes GPT-5.6 Sol.

Table 6: **Full evaluation with observed controls and uncertainty.** Accuracy (%) and visual gains Δ (percentage points), with pointwise 95% confidence intervals in brackets. Δ Independent compares against Observed vote@3; Δ Joint compares against Observed Joint control. Avg. weights datasets equally.

		Qwen3-VL			Kimi-K3			Gemini			GPT	
Space	Setting	8B	32B	235B	low	gw	max	2.5-Pro	3-Fl	3.7-Fl	5.2	5.6
Avg. (equal dataset weights)												
Observed	Single	50.0 [45.9, 54.3]	56.6 [52.2, 61.0]	55.9 [51.7, 60.2]	66.5 [62.3, 70.7]	67.4 [63.2, 71.5]	74.3 [70.3, 78.3]	63.3 [59.1, 67.4]	66.1 [62.3, 69.8]	75.5 [71.6, 79.2]	67.5 [63.1, 71.5]	70.1 [66.0, 74.1]
	vote@3	50.6 [46.4, 54.9]	56.5 [52.1, 60.9]	55.1 [50.9, 59.3]	69.0 [64.8, 73.0]	70.4 [66.3, 74.4]	75.0 [71.0, 78.9]	65.5 [61.3, 69.5]	66.2 [62.4, 69.8]	77.1 [73.4, 80.6]	68.7 [64.5, 72.7]	70.1 [66.0, 74.2]
	Joint control	47.7 [43.4, 51.9]	55.3 [51.0, 59.6]	51.5 [47.2, 55.6]	66.0 [61.8, 70.2]	66.9 [62.7, 71.0]	73.3 [69.2, 77.2]	61.6 [57.4, 65.8]	67.3 [63.4, 71.1]	76.4 [72.7, 80.0]	67.2 [63.1, 71.2]	69.8 [65.6, 73.8]
Language	Structured	50.7 [46.5, 55.0]	56.5 [52.3, 60.8]	55.9 [51.6, 60.2]	69.0 [64.8, 73.1]	67.9 [63.8, 72.0]	75.2 [71.4, 78.9]	62.0 [57.8, 66.1]	66.5 [62.8, 70.1]	76.8 [73.1, 80.2]	68.6 [64.2, 72.6]	71.7 [67.7, 75.6]
	Structured vote@3	50.6 [46.5, 54.9]	57.0 [52.7, 61.3]	54.8 [50.6, 59.1]	70.1 [65.9, 74.0]	69.9 [65.7, 74.0]	74.7 [70.7, 78.5]	64.9 [60.7, 69.0]	66.4 [62.6, 70.0]	76.3 [72.5, 79.9]	68.5 [64.2, 72.6]	70.7 [66.5, 74.7]
	Future	49.4 [45.3, 53.7]	58.2 [53.9, 62.3]	57.4 [53.1, 61.6]	68.5 [64.3, 72.6]	68.8 [64.7, 72.9]	75.3 [71.3, 79.2]	60.9 [56.5, 65.2]	65.1 [61.1, 68.9]	76.6 [72.8, 80.1]	66.3 [61.8, 70.4]	69.3 [65.2, 73.4]
	Future vote@3	49.5 [45.4, 53.7]	58.0 [53.7, 62.2]	56.4 [52.2, 60.8]	68.9 [64.7, 72.9]	71.2 [67.0, 75.2]	75.0 [71.0, 79.0]	63.1 [58.8, 67.2]	65.3 [61.4, 69.0]	78.2 [74.5, 81.6]	67.1 [62.7, 71.3]	68.8 [64.6, 72.9]
Visual	Independent	62.2 [57.9, 66.5]	69.5 [65.4, 73.6]	70.2 [66.1, 74.2]	76.3 [72.7, 79.7]	74.8 [70.9, 78.5]	76.4 [72.6, 80.0]	74.0 [70.2, 77.6]	74.7 [70.8, 78.2]	73.1 [69.2, 76.9]	72.0 [67.9, 75.9]	75.2 [71.3, 78.9]
	Joint	50.6 [46.3, 54.9]	65.8 [61.6, 69.7]	65.7 [61.5, 69.9]	76.2 [72.3, 79.9]	77.5 [73.6, 81.1]	79.2 [75.7, 82.6]	73.2 [69.4, 76.9]	77.0 [73.4, 80.5]	82.5 [79.5, 85.3]	75.1 [71.2, 78.9]	76.7 [72.9, 80.3]
Gain	Δ Independent	+11.6 [7.2, 16.2]	+13.0 [7.9, 18.1]	+15.1 [9.9, 20.2]	+7.4 [2.4, 12.4]	+4.4 [-0.9, 9.7]	+1.4 [-3.8, 6.4]	+8.5 [3.3, 13.8]	+8.4 [3.1, 13.7]	-3.9 [-9.1, 1.2]	+3.3 [-2.3, 8.8]	+5.1 [-0.3, 10.3]
Gain	Δ Joint	+2.9 [-0.0, 5.9]	+10.5 [5.9, 15.2]	+14.2 [9.7, 18.8]	+10.2 [5.4, 14.9]	+10.6 [5.6, 15.6]	+5.9 [1.1, 10.7]	+11.6 [6.9, 16.5]	+9.7 [4.7, 14.9]	+6.1 [1.1, 11.0]	+7.9 [3.0, 12.7]	+6.9 [1.9, 11.9]
Physion ($n = 1{,}035$)												
Observed	Single	47.4 [44.4, 50.4]	52.6 [49.5, 55.6]	51.1 [48.2, 54.1]	56.7 [53.7, 59.7]	56.7 [53.7, 59.7]	67.2 [64.4, 70.0]	55.7 [52.6, 58.7]	55.9 [52.9, 58.9]	66.1 [63.2, 69.0]	60.3 [57.3, 63.3]	62.4 [59.4, 65.3]
	vote@3	47.4 [44.4, 50.4]	52.6 [49.5, 55.6]	51.2 [48.3, 54.2]	58.1 [55.1, 61.1]	58.9 [55.9, 62.0]	66.7 [63.8, 69.6]	58.1 [55.2, 61.1]	55.9 [52.9, 58.9]	66.2 [63.3, 69.1]	60.9 [57.9, 63.9]	62.7 [59.8, 65.6]
	Joint control	49.5 [46.5, 52.5]	49.6 [46.5, 52.7]	48.1 [45.1, 51.1]	55.2 [52.2, 58.2]	53.7 [50.7, 56.8]	63.8 [60.9, 66.7]	58.0 [55.0, 61.1]	59.4 [56.3, 62.4]	65.9 [63.1, 68.8]	61.4 [58.5, 64.4]	62.9 [60.0, 65.8]
Language	Structured	49.2 [46.1, 52.2]	50.8 [47.8, 53.9]	50.6 [47.6, 53.6]	57.3 [54.3, 60.3]	56.1 [53.1, 59.1]	64.6 [61.7, 67.5]	56.4 [53.4, 59.5]	55.5 [52.4, 58.5]	64.7 [61.9, 67.6]	61.8 [58.8, 64.8]	62.5 [59.6, 65.4]
	Structured vote@3	49.1 [46.1, 52.1]	51.0 [47.9, 54.0]	50.6 [47.6, 53.6]	57.8 [54.9, 60.9]	58.7 [55.7, 61.8]	64.5 [61.7, 67.4]	56.5 [53.5, 59.5]	55.3 [52.2, 58.3]	66.5 [63.7, 69.4]	62.6 [59.6, 65.5]	64.5 [61.6, 67.3]
	Future	48.5 [45.5, 51.5]	54.0 [51.0, 57.0]	52.5 [49.6, 55.5]	58.3 [55.3, 61.4]	58.1 [55.1, 61.2]	67.5 [64.7, 70.4]	56.3 [53.2, 59.3]	56.1 [53.1, 59.0]	65.4 [62.6, 68.3]	62.1 [59.2, 65.1]	60.4 [57.4, 63.3]
	Future vote@3	48.5 [45.5, 51.5]	54.0 [51.0, 57.0]	52.6 [49.7, 55.6]	58.1 [55.1, 61.1]	61.8 [58.9, 64.8]	67.5 [64.7, 70.4]	56.7 [53.7, 59.7]	55.4 [52.3, 58.4]	67.0 [64.2, 69.8]	61.5 [58.6, 64.5]	61.4 [58.5, 64.3]
Visual	Independent	56.4 [53.4, 59.4]	60.9 [57.9, 63.8]	60.4 [57.4, 63.4]	59.3 [56.3, 62.3]	59.9 [56.9, 62.9]	61.6 [58.6, 64.5]	60.1 [57.0, 63.1]	56.8 [53.7, 59.8]	58.6 [55.6, 61.5]	60.5 [57.5, 63.5]	60.9 [57.9, 63.9]
	Joint	48.9 [45.9, 51.9]	60.3 [57.4, 63.3]	62.6 [59.7, 65.7]	62.6 [59.6, 65.5]	64.4 [61.5, 67.3]	63.7 [60.8, 66.6]	59.7 [56.7, 62.6]	64.4 [61.5, 67.3]	64.0 [61.1, 67.0]	63.0 [60.0, 66.0]	61.6 [58.6, 64.5]
Gain	Δ Independent	+9.0 [5.4, 12.6]	+8.3 [4.6, 11.9]	+9.2 [5.7, 12.7]	+1.3 [-2.2, 4.6]	+1.0 [-2.7, 4.4]	-5.0 [-8.7, -1.4]	+2.0 [-2.0, 6.1]	+0.9 [-3.4, 5.1]	-7.6 [-12.0, -3.3]	-0.4 [-4.8, 3.8]	-1.8 [-6.3, 2.7]
Gain	Δ Joint	-0.6 [-2.7, 1.5]	+10.7 [7.0, 14.6]	+14.5 [10.7, 18.3]	+7.4 [4.0, 10.8]	+10.7 [7.2, 14.3]	-0.1 [-3.5, 3.3]	+1.7 [-2.3, 5.7]	+5.0 [0.9, 9.3]	-1.9 [-5.8, 2.0]	+1.5 [-2.3, 5.2]	-1.3 [-5.4, 2.8]

Table 6: **Full evaluation with observed controls and uncertainty (continued).** Accuracy (%) and visual gains Δ (percentage points), with pointwise 95% confidence intervals in brackets. Δ Independent compares against Observed vote@3; Δ Joint compares against Observed Joint control.

		Qwen3-VL			Kimi-K3			Gemini			GPT	
Space	Setting	8B	32B	235B	low	gw	max	2.5-Pro	3-Fl	3.7-Fl	5.2	5.6
CLEVRER ($n = 1{,}000$)												
Observed	Single	63.2 [60.2, 66.2]	65.8 [62.9, 68.7]	68.1 [65.2, 71.0]	86.7 [84.6, 88.7]	86.4 [84.2, 88.5]	89.1 [87.2, 91.0]	72.2 [69.5, 74.9]	66.7 [63.8, 69.6]	89.1 [87.1, 91.0]	80.1 [77.6, 82.5]	85.8 [83.6, 87.9]
	vote@3	63.4 [60.4, 66.4]	65.5 [62.5, 68.4]	68.6 [65.8, 71.5]	86.7 [84.6, 88.7]	87.0 [84.9, 89.0]	90.1 [88.2, 91.9]	74.7 [72.0, 77.3]	67.0 [64.2, 69.9]	89.3 [87.3, 91.1]	81.5 [79.1, 83.8]	85.5 [83.3, 87.6]
	Joint control	51.3 [48.2, 54.4]	63.3 [60.3, 66.2]	62.3 [59.3, 65.4]	85.4 [83.2, 87.5]	83.2 [80.9, 85.4]	88.0 [85.9, 90.0]	70.8 [68.1, 73.6]	71.3 [68.5, 74.0]	89.1 [87.1, 91.0]	78.1 [75.5, 80.6]	84.3 [82.0, 86.5]
Language	Structured	63.4 [60.4, 66.4]	67.2 [64.3, 70.1]	68.5 [65.6, 71.4]	86.0 [83.8, 88.1]	86.9 [84.7, 88.9]	88.3 [86.3, 90.2]	71.9 [69.0, 74.6]	66.9 [64.0, 69.8]	88.4 [86.4, 90.3]	80.2 [77.7, 82.6]	84.3 [82.0, 86.5]
	Structured vote@3	63.2 [60.1, 66.2]	67.1 [64.2, 70.1]	68.4 [65.5, 71.3]	85.9 [83.7, 88.0]	87.3 [85.2, 89.3]	89.8 [87.9, 91.6]	74.5 [71.8, 77.1]	68.2 [65.3, 71.1]	89.8 [87.9, 91.6]	82.3 [79.9, 84.6]	85.3 [83.0, 87.5]
	Future	63.3 [60.2, 66.3]	65.9 [62.9, 68.8]	68.2 [65.3, 71.1]	86.6 [84.5, 88.7]	86.3 [84.1, 88.4]	88.7 [86.7, 90.6]	73.4 [70.6, 76.1]	66.3 [63.5, 69.2]	90.1 [88.2, 91.9]	80.8 [78.3, 83.2]	85.4 [83.2, 87.5]
	Future vote@3	63.5 [60.5, 66.5]	66.9 [64.0, 69.8]	68.3 [65.4, 71.1]	86.4 [84.3, 88.5]	86.5 [84.3, 88.6]	89.3 [87.4, 91.2]	74.9 [72.2, 77.6]	67.7 [64.8, 70.6]	90.4 [88.5, 92.2]	80.7 [78.2, 83.1]	85.9 [83.7, 88.0]
Visual	Independent	74.2 [71.5, 77.0]	85.5 [83.3, 87.6]	88.2 [86.2, 90.2]	90.9 [89.1, 92.7]	93.3 [91.7, 94.8]	94.8 [93.4, 96.1]	89.1 [87.1, 91.0]	92.9 [91.3, 94.5]	92.7 [91.0, 94.3]	90.4 [88.6, 92.2]	93.4 [91.8, 94.9]
	Joint	57.5 [54.4, 60.5]	73.4 [70.6, 76.2]	78.3 [75.7, 80.8]	94.8 [93.4, 96.1]	93.7 [92.2, 95.2]	96.7 [95.6, 97.8]	85.6 [83.4, 87.8]	89.4 [87.5, 91.3]	97.2 [96.1, 98.2]	94.2 [92.7, 95.6]	94.2 [92.7, 95.6]
Gain	Δ Independent	+10.8 [7.9, 13.7]	+20.0 [17.0, 23.0]	+19.6 [16.8, 22.4]	+4.2 [1.7, 6.8]	+6.3 [4.0, 8.7]	+4.7 [2.5, 6.9]	+14.4 [11.4, 17.3]	+25.9 [22.8, 28.9]	+3.4 [1.2, 5.6]	+8.9 [6.1, 11.7]	+7.9 [5.5, 10.3]
Gain	Δ Joint	+6.2 [3.7, 8.7]	+10.1 [7.2, 12.9]	+16.0 [13.0, 19.0]	+9.4 [7.4, 11.5]	+10.5 [8.3, 12.8]	+8.7 [6.7, 10.7]	+14.8 [11.6, 18.0]	+18.1 [15.0, 21.2]	+8.1 [6.2, 10.0]	+16.1 [13.8, 18.6]	+9.9 [7.8, 12.1]
Physics-IQ ($n = 66$)												
Observed	Single	39.4 [27.3, 51.5]	51.5 [39.4, 63.6]	48.5 [36.4, 60.6]	56.1 [43.9, 68.2]	59.1 [47.0, 71.2]	66.7 [54.5, 77.3]	62.1 [50.0, 74.2]	75.8 [65.2, 86.4]	71.2 [60.6, 81.8]	62.1 [50.0, 72.8]	62.1 [50.0, 74.2]
	vote@3	40.9 [28.8, 53.0]	51.5 [39.4, 63.6]	45.5 [33.3, 57.6]	62.1 [50.0, 74.2]	65.2 [53.0, 75.8]	68.2 [56.1, 78.8]	63.6 [51.5, 75.8]	75.8 [65.2, 86.4]	75.8 [65.2, 84.8]	63.6 [51.5, 74.2]	62.1 [50.0, 74.2]
	Joint control	42.4 [30.3, 54.5]	53.0 [40.9, 65.2]	43.9 [31.8, 56.1]	57.6 [45.5, 69.7]	63.6 [51.5, 75.8]	68.2 [56.1, 78.8]	56.1 [43.9, 68.2]	71.2 [60.6, 81.8]	74.2 [63.6, 84.8]	62.1 [50.0, 72.7]	62.1 [50.0, 74.2]
Language	Structured	39.4 [27.3, 51.5]	51.5 [39.4, 63.6]	48.5 [36.4, 60.6]	63.6 [51.5, 75.8]	60.6 [48.5, 72.7]	72.7 [62.1, 83.3]	57.6 [45.5, 69.7]	77.3 [66.7, 87.9]	77.3 [66.7, 86.4]	63.6 [51.5, 75.8]	68.2 [57.6, 78.8]
	Structured vote@3	39.4 [27.3, 51.5]	53.0 [40.9, 65.2]	45.5 [33.3, 57.6]	66.7 [54.5, 77.3]	63.6 [51.5, 75.8]	69.7 [59.1, 80.3]	63.6 [51.5, 75.8]	75.8 [65.2, 86.4]	72.7 [62.1, 83.3]	60.6 [48.5, 72.7]	62.1 [50.0, 74.2]
	Future	36.4 [24.2, 48.5]	54.5 [42.4, 66.7]	51.5 [39.4, 63.6]	60.6 [48.5, 72.7]	62.1 [50.0, 74.2]	69.7 [57.6, 80.3]	53.0 [40.9, 65.2]	72.7 [62.1, 83.3]	74.2 [63.6, 84.8]	56.1 [43.9, 68.2]	62.1 [50.0, 74.2]
	Future vote@3	36.4 [24.2, 48.5]	53.0 [40.9, 65.2]	48.5 [36.4, 60.6]	62.1 [50.0, 74.2]	65.2 [53.0, 75.8]	68.2 [56.1, 78.8]	57.6 [45.5, 69.7]	72.7 [62.1, 83.3]	77.3 [66.7, 86.4]	59.1 [47.0, 71.2]	59.1 [47.0, 71.2]
Visual	Independent	56.1 [43.9, 68.2]	62.1 [50.0, 74.2]	62.1 [50.0, 74.2]	78.8 [68.2, 87.9]	71.2 [60.6, 81.8]	72.7 [62.1, 83.3]	72.7 [62.1, 83.3]	74.2 [63.6, 84.8]	68.2 [57.6, 78.8]	65.2 [53.0, 75.8]	71.2 [60.6, 81.8]
	Joint	45.5 [33.3, 57.6]	63.6 [51.5, 74.2]	56.1 [43.9, 68.2]	71.2 [60.6, 81.8]	74.2 [63.6, 84.8]	77.3 [66.7, 86.4]	74.2 [63.6, 84.8]	77.3 [66.7, 86.4]	86.4 [77.3, 93.9]	68.2 [56.1, 78.8]	74.2 [63.6, 84.8]
Gain	Δ Independent	+15.2 [3.0, 27.3]	+10.6 [-4.5, 25.8]	+16.7 [1.5, 31.8]	+16.7 [3.0, 30.3]	+6.1 [-9.1, 21.2]	+4.5 [-10.6, 18.2]	+9.1 [-6.1, 24.2]	-1.5 [-16.7, 13.6]	-7.6 [-22.7, 7.6]	+1.5 [-15.2, 16.7]	+9.1 [-6.1, 24.2]
Gain	Δ Joint	+3.0 [-4.5, 12.1]	+10.6 [-3.0, 24.2]	+12.1 [0.0, 25.8]	+13.6 [0.0, 27.3]	+10.6 [-3.0, 24.2]	+9.1 [-4.5, 22.7]	+18.2 [4.5, 31.8]	+6.1 [-7.6, 19.7]	+12.1 [-1.5, 25.8]	+6.1 [-7.6, 19.7]	+12.1 [-1.6, 25.8]

C Additional Qualitative Results

C.1 Webpage with additional videos and examples

The project webpage provides a more complete qualitative view than can be shown in the PDF. It includes additional success and failure cases, the observed prefixes, multiple generated rollouts, and model predictions across the three benchmarks. We use the PDF supplement for a small number of representative failure modes and use the webpage for broader visual coverage.

Additional videos and examples are available on the project webpage.

C.2 Failure cases and limitations

The main qualitative figure emphasizes cases where imagined futures help. Figure 4 complements it with two representative failure modes: one in which the VLM misinterprets useful visual evidence and one in which the generated continuation itself is misleading.

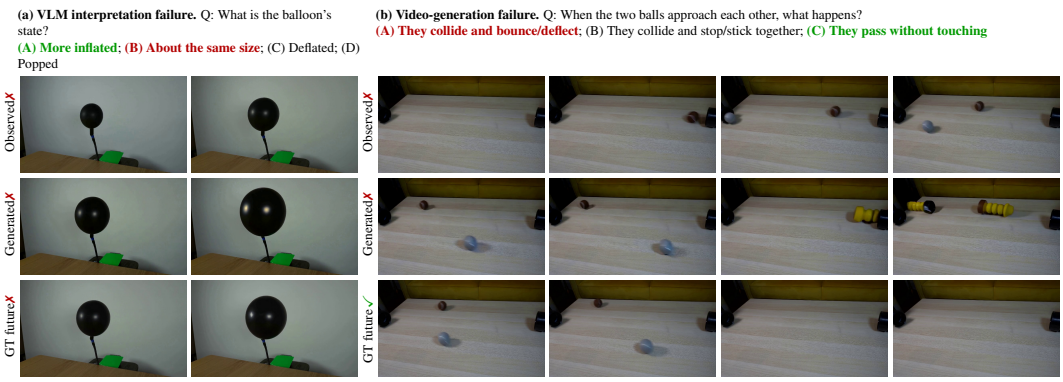


Figure 4: **Representative failure modes.** (a) The imagined and ground-truth futures both show the balloon growing, but the VLM still predicts that its size is approximately unchanged, illustrating a downstream visual-interpretation error. (b) The generated continuation diverges from the physical outcome and introduces misleading evidence, while the ground-truth future shows the two balls passing without contact.

These examples expose two distinct bottlenecks. Better generators can reduce the second failure mode, but they do not by themselves solve the first: even when the relevant future event is visible, the VLM must still identify the objects, track the state change, and map that evidence to the answer. This separation is also why we treat the GT-video condition as a diagnostic reference rather than as a guaranteed upper bound.