VideoHallu: Evaluating and Mitigating Multi-modal Hallucinations on Synthetic Video Understanding

Zongxia Li[†][*] **Xiyang Wu**[†][*] **Guangyao Shi**[‡] **Yubin Qin**[†] **Hongyang Du**[†]
Tianyi Zhou[†] **Dinesh Manocha**[†] **Jordan Lee Boyd-Graber**[†]

[†]University of Maryland, College Park [‡]University of Southern California

{zli12321, wuxiyang, Yubinq, hydu, zhou, dmanocha, ying}@umd.edu,
shig@usc.edu

Abstract

Vision–Language Models (VLMs) have achieved remarkable success in video understanding tasks. Yet, a key question remains: do they comprehend visual information, or merely learn superficial mappings between visual and textual patterns? Understanding visual cues, particularly those related to physics and common sense, is crucial for AI systems interacting with the physical world. However, existing VLM evaluations primarily rely on *positive-control* tests using real-world videos that resemble training distributions. While VLMs perform well on such benchmarks, it is unclear whether they grasp underlying visual and contextual signals or simply exploit visual-language correlations. To fill this gap, we propose incorporating *negative-control* tests, *i.e.*, videos depicting physically impossible or logically inconsistent scenarios, and evaluating whether models can recognize these violations. True visual understanding should evince comparable performance across both positive and negative tests. Since such content is rare in the real world, we introduce VideoHallu, a synthetic video dataset featuring physics- and commonsense-violating scenes generated using state-of-the-art tools such as Veo2, Sora, and Kling. The dataset includes expert-annotated question–answer pairs spanning four categories of physical and commonsense violations, designed to be straightforward for human reasoning. We evaluate several leading VLMs, including Qwen-2.5-VL, Video-R1, and VideoChat-R1. Despite their strong performance on real-world benchmarks (*e.g.*, MVBench, MMVU), these models hallucinate or fail to detect physical or logical violations, revealing fundamental weaknesses in visual understanding. Finally, we explore reinforcement learning–based post-training on our *negative* dataset: fine-tuning improves performance on VideoHallu without degrading results on standard benchmarks—indicating enhanced visual reasoning in VLMs. Our data is available at https://github.com/zli12321/VideoHallu.git.

1 Introduction

Vision–Language Models (VLMs) have made remarkable progress in video understanding. However, they remain prone to hallucinations and shallow visual reasoning [1–3]. Prior works mitigate these issues across various domains, including chart interpretation [4], video understanding [5], and visual question answering (VQA) [6], primarily through supervised fine-tuning (SFT) or R1-style chain-of-thought training (reinforcement learning) [7, 8]. However, most of these VLM evaluations rely on *positive-control* test, that is, real-world data drawn from distributions closely aligned with training data. Consequently, it remains unclear whether current VLMs genuinely reason about visual cues or merely exploit prior visual-language correlations within familiar distributions [9].

To truly evaluate visual understanding, mwe test VLMs under *negative-control* conditions, *i.e.*, videos outside their training distribution that depict physically impossible or logically inconsistent events.

[*] Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

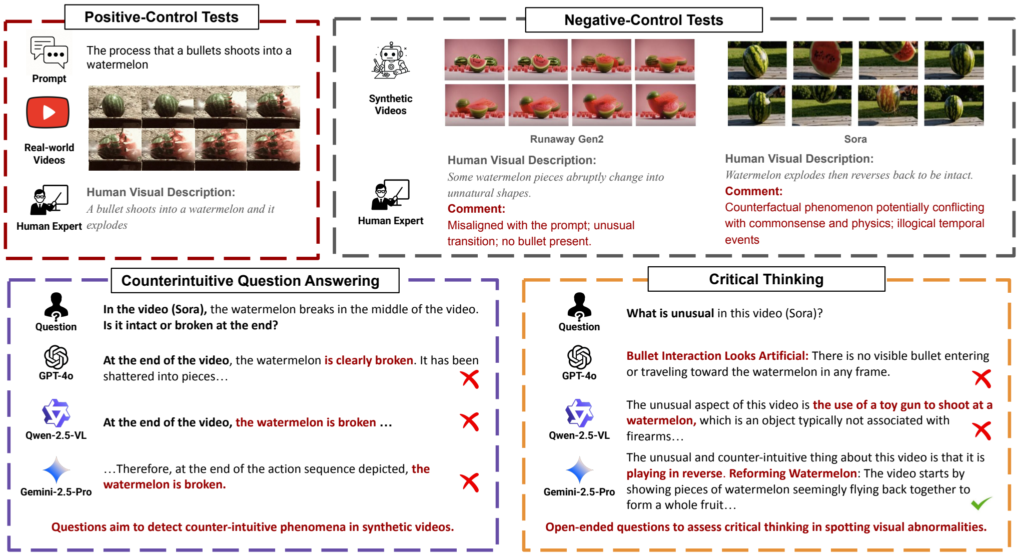

Positive-Control Tests
Prompt
The process that a bullets shoots into a watermelon
Real-world Videos
Human Expert
Human Visual Description:
A bullet shoots into a watermelon and it explodes
Negative-Control Tests
Synthetic Videos
Runaway Gen2
Sora
Human Expert
Human Visual Description:
Some watermelon pieces abruptly change into unnatural shapes.
Comment:
Misaligned with the prompt; unusual transition; no bullet present.
Human Visual Description:
Watermelon explodes then reverses back to be intact.
Comment:
Counterfactual phenomenon potentially conflicting with commonsense and physics; illogical temporal events
Counterintuitive Question Answering
Question
In the video (Sora), the watermelon breaks in the middle of the video. Is it intact or broken at the end?
GPT-4o
At the end of the video, the watermelon is clearly broken. It has been shattered into pieces…
Qwen-2.5-VL
At the end of the video, the watermelon is broken …
Gemini-2.5-Pro
…Therefore, at the end of the action sequence depicted, the watermelon is broken.
Questions aim to detect counter-intuitive phenomena in synthetic videos.
Critical Thinking
Question
What is unusual in this video (Sora)?
GPT-4o
Bullet Interaction Looks Artificial: There is no visible bullet entering or traveling toward the watermelon in any frame.
Qwen-2.5-VL
The unusual aspect of this video is the use of a toy gun to shoot at a watermelon, which is an object typically not associated with firearms…
Gemini-2.5-Pro
The unusual and counter-intuitive thing about this video is that it is playing in reverse. Reforming Watermelon: The video starts by showing pieces of watermelon seemingly flying back together to form a whole fruit…
Open-ended questions to assess critical thinking in spotting visual abnormalities.

Figure 1: **Illustrative examples of designed negative-control tests to evaluate the critical thinking abilities of VLMs.** Unlike real-world videos, synthetic videos can contain counterfactual or commonsense-violating contexts misaligned with reality. VideoHallu includes such synthetic videos with perceptually obvious abnormalities, paired with crafted questions that probe counterintuitive phenomena or test VLMs' critical thinking in detecting such abnormalities. When SOTA VLMs are evaluated on VideoHallu, they frequently hallucinate, which suggests that these models rely on language priors and commonsense knowledge rather than truly understand the videos.

These tests reveal whether models detect violations of physics, causality, or commonsense instead of relying on memorized language knowledge. However, constructing such out-of-distribution (OOD) videos in the real world is costly and impractical [10].

Modern video generation models such as Veo2, Sora, and Runway [11–14] can produce photorealistic but physically impossible scenes. Such models provide an alternative option to generate test videos for probing VLMs' visual understanding. By careful design, these synthetic videos can be systematically introduced to include violations of gravity, causality, and commonsense interactions [15, 16], enabling controlled OOD evaluations where models need to rely on visual cues. Current VLMs, predominantly trained on videos conforming to physical laws, may thus overfit to statistical regularities rather than learning genuine causal reasoning [17]. Figure 1 illustrates that even state-of-the-art VLMs such as Gemini-2.5-Pro [18], GPT-4o [19], and Qwen2.5-VL [20] hallucinate when confronted with counterintuitive scenarios. For instance, when a watermelon reassembles after an explosion, models rely on linguistic priors (*e.g.*, *"a watermelon should break when shot"*) rather than actual visual cues, exposing their limited physical reasoning.

To rigorously evaluate such limitations, we introduce VideoHallu, a dataset of expert-curated question–answer pairs over synthetic videos featuring controlled violations of alignment, spatial–temporal consistency, commonsense, and physics. We benchmark several leading VLMs and analyze their failure modes on these physically and logically inconsistent videos. Finally, we explore two post-training strategies, supervised fine-tuning (SFT) and reinforcement learning (RL) via Group Relative Policy Optimization (GRPO) [21], using both real-world data from Video-R1 [7] and synthetic data from VideoHallu. GRPO enhances generalization on synthetic video reasoning without degrading real-world performance.

Contributions.

1) We introduce VideoHallu, a dataset of 3K expert-annotated QA pairs on synthetic videos that include violations spanning alignment, consistency, commonsense, and physical reasoning.

2) We evaluate state-of-the-art VLMs and find that even top-performing models (*e.g.*, GPT-4o, Gemini-2.5-Pro) can achieve only ∼50% accuracy in our dataset, exhibiting frequent hallucination in counterintuitive scenarios.

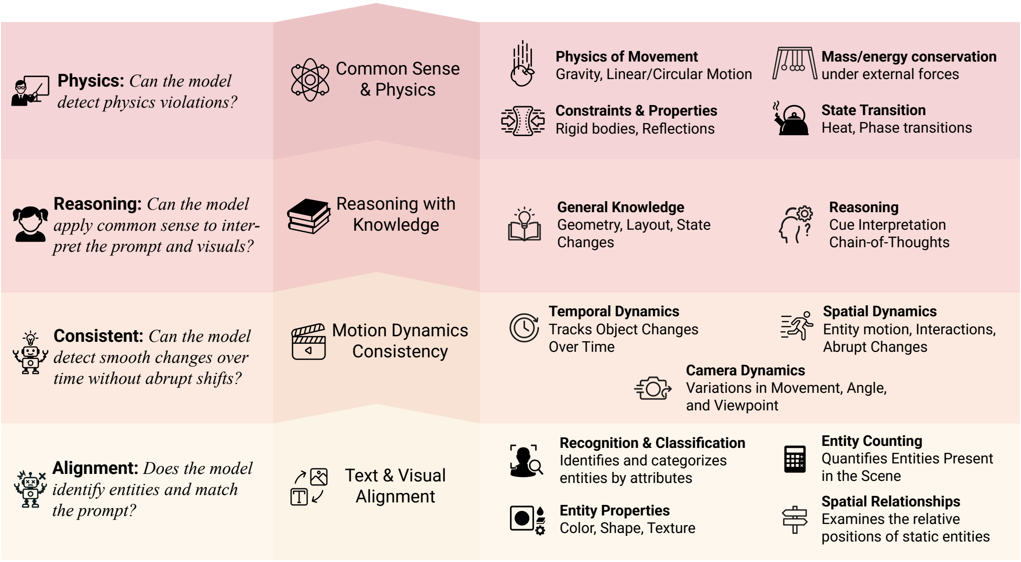

Physics: Can the model detect physics violations?
Common Sense & Physics
Physics of Movement
Gravity, Linear/Circular Motion
Mass/energy conservation under external forces
Constraints & Properties
Rigid bodies, Reflections
State Transition
Heat, Phase transitions
Reasoning: Can the model apply common sense to interpret the prompt and visuals?
Reasoning with Knowledge
General Knowledge
Geometry, Layout, State Changes
Reasoning
Cue Interpretation
Chain-of-Thoughts
Consistent: Can the model detect smooth changes over time without abrupt shifts?
Motion Dynamics Consistency
Temporal Dynamics
Tracks Object Changes Over Time
Spatial Dynamics
Entity motion, Interactions, Abrupt Changes
Camera Dynamics
Variations in Movement, Angle, and Viewpoint
Alignment: Does the model identify entities and match the prompt?
Text & Visual Alignment
Recognition & Classification
Identifies and categorizes entities by attributes
Entity Counting
Quantifies Entities Present in the Scene
Entity Properties
Color, Shape, Texture
Spatial Relationships
Examines the relative positions of static entities

Figure 2: **Question Categorization of VideoHallu.** We design our benchmark, VideoHallu, with four question categories to probe limitations in synthetic video understanding, covering perceptual understanding to abstract reasoning: **(a) Physics** assesses if the model applies physical laws to entity motions and procedural understanding. **(b) Common Sense Reasoning** tests if the model can reason based on its knowledge. **(c) Spatial-temporal Consistency** examines whether the model can track entity motion across frames. **(d) Alignment** checks if the model correctly identifies and understands entities using visual and textual cues.

3) GRPO post-training with synthetic data improves visual reasoning on VideoHallu while maintaining real-world performance, providing a path toward more physically grounded VLMs.

2 VideoHallu: Evaluating VLMs' Synthetic Video Understanding

Preliminary. Our objective is to evaluate whether VLMs can effectively reason about and answer questions concerning synthetic videos that fall outside the distribution of their training data. To construct such evaluation data, we synthesize videos from text prompts using generative models. For these synthetic test videos to be meaningful, they should incorporate deliberate visual abnormalities that elicit responses contradicting common-sense expectations—situations in which a model relying solely on linguistic reasoning would produce one answer, but where careful visual observation reveals an alternative, visually grounded truth. To see how VLMs handle synthetic videos with such *abnormalities*, we categorize our evaluation questions into *counter-intuitive* and *critical thinking* types. Counter-intuitive questions focus on implausible or physically impossible events (*e.g.*, a shattered watermelon reassembling itself), while critical thinking questions evaluate the model's ability to detect visual inconsistencies or logical contradictions (*e.g.*, unnatural object breakage).

Data Collection. Our data collection pipeline consists of two main stages: The first stage generates synthetic videos V with common sense or physics abnormalities, *i.e.*, videos that satisfy constraints (5), where the LLM backbone possesses human-aligned knowledge but VLMs overlook abnormalities, resulting in answers that disagree with human perception. We recruited five human experts to review the defined abnormality categories (detailed in Table 2 and Appendix C) and craft prompts that reproduce such abnormalities in generated synthetic videos. In total, they created 141 adversarial prompts, used to generate 987 videos across seven models: Sora [22], Veo2 [11], Runway Gen 2 [13], Kling [23], Pixverse [24], Lavie [25], and CogVideo [26].

In the second stage, we craft adversarial video QA pairs to evaluate VLMs' understanding of synthetic videos. Human experts manually review each video to identify counterintuitive contexts that lead to significant discrepancies between VLM outputs and human perception, *i.e.*, video QA pairs maximizing the objective function (4). They then construct natural language questions—along with the ground truth answer—based on the context. These QA pairs are categorized into sub-categories (Table 2). Each annotator writes QA pairs highlighting visually clear but semantically abnormal content, difficult for VLMs to detect. These questions are not designed to trick models but rather to probe their ability to catch subtle violations of common sense, physics, or prompt-video mismatches, critical for robust, interpretable video evaluation (Figure 3).

Dataset Metadata. Our dataset comprises 3,233 video question–answer pairs with no video overlap across splits: 800 pairs for training, 908 for validation, and 1,525 for testing. The videos average 96.0 frames per video corresponding to approximately 5.3 seconds at an average framerate of 23 FPS. Frame resolution averages 1042 × 588 pixels across all videos.

3 Experiment and Results

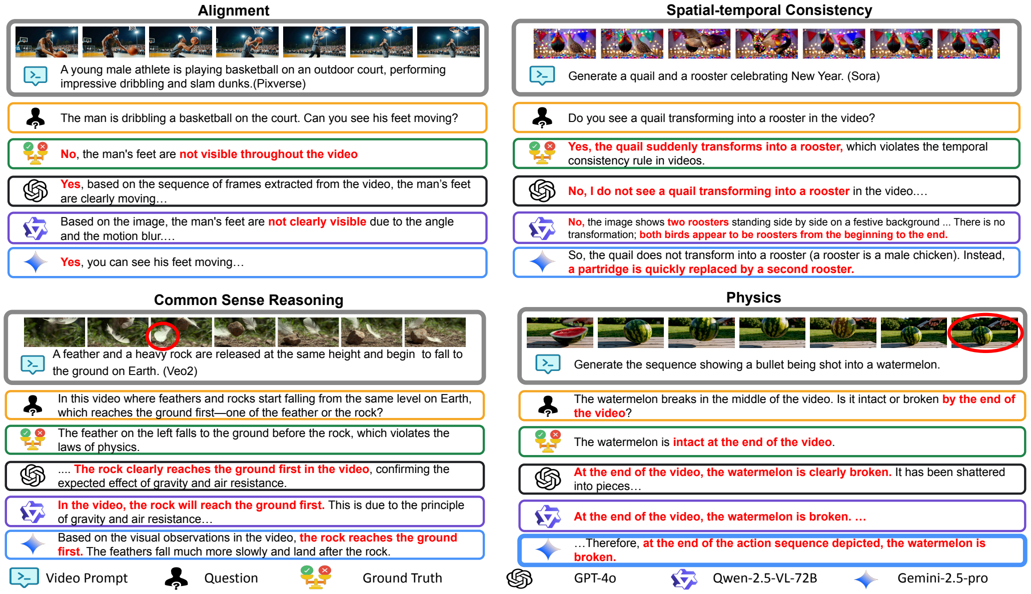
Given the collected adversarial QA pairs, we evaluate 17 SOTA VLMs (Table 1). For models not trained with RL or chain-of-thought (CoT) generation, we use standard prompting to generate direct answers. For those trained with RL or CoT supervised finetuning (*e.g.*, Video-R1-CoT [7] and VideoChat-R1-think [8]), we prompt them to generate step-by-step critical thinking and reasoning before generating a final answer (Appendix. D). Figure 3 highlights hallucinations produced by SoTA models across all four categories in synthetic video understanding tasks, with the hallucinated contexts marked within each answer (additional examples in Appendix A).

Answer Evaluation: We adopt LLM-as-a-Judge [27–29] as our evaluation method. GPT-4o-mini evaluates the correctness of model responses (§ 6).[2]

Model	Alignment	Physics	Consistency	Commonsense	Overall
VLMs: <7B					
`SmolVLM-3B` [30]	15.94	13.44	22.49	8.75	16.13
`Qwen2.5-VL-3B` [20]	41.53	27.21	26.91	26.25	**35.48**
`InternVL3-2B` [31]	47.36	32.79	42.17	32.50	42.82
VLMs: >7B					
`LLaVA-OneVision` [32]	44.22	32.46	32.13	45.00	39.93
`Video-LLaVA` [33]	46.58	40.00	43.37	31.25	43.93
`LLaVA-NeXT` [34]	50.95	36.07	38.96	31.25	44.98
`Video-LLaMA` [35]	55.67	38.69	50.20	32.50	50.16
`InternVL3-9B` [31]	53.54	43.61	47.79	38.75	49.84
`InternVL3-14B` [31]	53.65	45.90	46.18	31.25	49.70
`InternVL3-38B` [31]	55.78	38.69	50.20	38.75	50.56
`Qwen2.5-VL-32B` [20]	58.81	42.95	46.59	40.00	52.66
`Qwen2.5-VL-7B` [20]	58.02	44.59	46.99	47.50	**52.98**
VLMs: R1-finetuned					
`VideoChat-R1` [8]	53.31	40.33	44.58	45.00	48.85
`Video-R1-SFT` [7]	58.14	47.21	48.19	41.25	53.44
`Video-R1` [7]	58.14	**48.20**	**49.00**	38.75	**53.64**
VLMs: Close-Source					
`Gemini-2.5-Pro` [18]	58.36	33.11	40.16	36.25	49.18
`Gemini-2.0-Flash` [36]	56.57	39.02	42.97	40.00	49.97
`GPT-4o-mini` [19]	54.88	41.97	48.19	38.75	**50.36**

Table 1: **Video Model Evaluation Results.** We evaluate diverse VLMs across different sizes on our test set, reporting alignment, physics understanding, spatial-temporal consistency, and commonsense reasoning. Video-R1 is best overall.

[2] For CoT generations, we extract the final answer as responses to evaluate. To validate reliability on our dataset, we manually annotated 200 randomly sampled answer pairs, achieving 97% agreement with GPT-4o-mini.



Alignment

A young male athlete is playing basketball on an outdoor court, performing impressive dribbling and slam dunks.(Pixverse)

The man is dribbling a basketball on the court. Can you see his feet moving?

No, the man's feet are **not visible throughout the video**

Yes, based on the sequence of frames extracted from the video, the man’s feet are clearly moving…

Based on the image, the man's feet are **not clearly visible** due to the angle and the motion blur.…

Yes, you can see his feet moving…

Spatial-temporal Consistency

Generate a quail and a rooster celebrating New Year. (Sora)

Do you see a quail transforming into a rooster in the video?

Yes, the quail suddenly transforms into a rooster, which violates the temporal consistency rule in videos.

No, I do not see a quail transforming into a rooster in the video.…

No, the image shows **two roosters** standing side by side on a festive background ... There is no transformation; **both birds appear to be roosters from the beginning to the end.**

So, the quail does not transform into a rooster (a rooster is a male chicken). Instead, **a partridge is quickly replaced by a second rooster.**

Common Sense Reasoning

A feather and a heavy rock are released at the same height and begin to fall to the ground on Earth. (Veo2)

In this video where feathers and rocks start falling from the same level on Earth, which reaches the ground first—one of the feather or the rock?

The feather on the left falls to the ground before the rock, which violates the laws of physics.

.... **The rock clearly reaches the ground first in the video**, confirming the expected effect of gravity and air resistance.

In the video, the rock will reach the ground first. This is due to the principle of gravity and air resistance…

Based on the visual observations in the video, **the rock reaches the ground first.** The feathers fall much more slowly and land after the rock.

Physics

Generate the sequence showing a bullet being shot into a watermelon.

The watermelon breaks in the middle of the video. Is it intact or broken **by the end of the video**?

The watermelon is **intact at the end of the video**.

At the end of the video, the watermelon is clearly broken. It has been shattered into pieces…

At the end of the video, the watermelon is broken. …

…Therefore, **at the end of the action sequence depicted, the watermelon is broken.**

Figure 3: **Example Synthetic Videos in VideoHallu.** Example hallucination cases observed during SOTA VLM evaluations on synthetic video tasks. Each example includes the generation prompt, key frames, questions, human-annotated ground truth, and hallucinated answers from GPT-4o, Qwen2.5-VL, and Gemini-2.5-Pro, with hallucinations marked in **Red**.

3.1 Limitations of VLMs in OOD Data

VLMs struggle with counterintuitive phenomena and abnormalities in generated videos. State-of-the-art VLMs achieve below 55% accuracy on our synthetic video QA dataset, only slightly above the 50% random baseline (Table 1). They particularly falter in commonsense and physical reasoning, often failing to detect abnormalities or relying on linguistic shortcuts instead of visual evidence. As shown in Figure 3, none of the models recognize the implausibility of a shattered watermelon reassembling, nor do they notice abrupt, counterfactual entity changes, such as a quail suddenly turning into a rooster. These failures highlight VLMs’ limited capacity for abnormality reasoning and critical visual understanding beyond text priors.

Chain-of-thought reasoning learned from real-world videos provides limited benefit for understanding synthetic videos. VLMs trained with reinforcement learning, such as GRPO [37] (R1-finetuned) used in the DeepSeek series [21], show potential for improving reasoning and critical thinking abilities in reasoning-heavy tasks like mathematics, real-world video understanding [38, 3, 39]. This raises a question: does RL truly improve visual reasoning in VLMs, or does it only optimize for correct answers without enhancing actual visual understanding? Table 1 evaluates two R1-finetuned models, Video-R1 [7] and VideoChat-R1 [8], using chain-of-thought prompts. Both models show limited improvement compared to their base models (Qwen2.5-VL-7B) on synthetic video understanding, with minimal or worse alignment and commonsense, suggesting that training on real-world videos only inculcates real-world reasoning patterns.

Solely pre-training on real-world data biases visual grounding. While RL improves reasoning on math and real-world videos, it does not help with counterintuitive synthetic content that contradicts real-world norms and cannot elicit video-grounded critical thinking. In such cases, chain-of-thought prompting can bias the LLM backbone to rely too heavily on prior commonsense knowledge, neglecting synthetic visual cues and leading to hallucinated responses [9]. For instance, in the third case in Figure 3, the video shows the feather dropping to the ground before the rock when falling from the same height. When asked which reaches the ground first—the feather or the rock, VideoChat-R1-think responds: *“The video shows a feather and a rock being dropped. . . This is a classic demonstration of Galileo’s principle; therefore the rock drops before the feather. . . ”* While this language explanation alone is grounded in correct physical principles, it directly contradicts what actually occurs in the video. The model generates an incorrect conclusion based on prior language reasoning, showing how chain-of-thought prompting can amplify reliance on language priors and increase hallucination risk when understanding synthetic videos that do not align with real-world expectations.

3.2 Visual Learning or Pattern Matching? Questioning RL's True Impact on Vision Models

Current VLMs struggle with counter-intuitive questions and critical visual thinking in synthetic videos (Section 3.1). They frequently hallucinate and a dearth of critical thinking leads them to overlook abnormality examples in VideoHallu, relying on the model's language knowledge instead of reasoning directly from the visual input. This raises a crucial question: *Can VLMs learn counter-intuitive commonsense knowledge and improve their critical thinking abilities for detecting abnormalities through training with synthetic video data?*

While supervised fine-tuning (SFT) or RL from human feedback (RLHF) are natural approaches for improving VLMs, the distributional gap between standard alignment-based video QA tasks used in pre-training and the specialized critical reasoning required for synthetic videos poses an impediment. Since synthetic videos are scarce in typical pre-training corpora, models lack sufficient exposure to develop robust reasoning capabilities for such content.

To investigate whether incorporating synthetic data alongside real data can improve VLM performance on synthetic videos, we pose two primary research questions:

- *Between SFT and GRPO training, which approach more effectively enables VLMs to develop a genuine understanding of synthetic videos?*
- *Is synthetic data in the training mixture necessary for improving model reasoning abilities on synthetic videos, or can training on real data alone suffice?*

Method Overview: SFT vs. GRPO. We compare two training paradigms—SFT and GRPO [21]—to compare their effectiveness and generalization.

Supervised Fine-Tuning (SFT). SFT directly optimizes the model to predict ground-truth responses using maximum likelihood estimation. For a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ containing video-question pairs x_i and corresponding answers y_i, the SFT objective minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{|y_i|} \log p_\theta \left(y_i^{(t)} \mid x_i, y_i^{(<t)} \right) \tag{1}$$

where θ represents model parameters, $y_i^{(t)}$ is the t-th token in sequence y_i, and $y_i^{(<t)}$ denotes all preceding tokens.

Group Relative Policy Optimization (GRPO). GRPO, a variant of reinforcement learning from human feedback, optimizes the model using preference-based learning without requiring explicit reward models. Given preference pairs (y_w, y_l) where y_w is preferred over y_l for prompt x, GRPO maximizes the likelihood of preferred responses while penalizing less preferred ones:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{p_\theta(y_w \mid x)}{p_{\text{ref}}(y_w \mid x)} - \beta \log \frac{p_\theta(y_l \mid x)}{p_{\text{ref}}(y_l \mid x)} \right) \right] \tag{2}$$

where p_{ref} is a reference model (typically the pre-trained checkpoint), β is a temperature parameter controlling the strength of the KL penalty, and σ is the sigmoid function. This approach encourages the model to generate responses that align better with the training data distribution while maintaining proximity to the reference policy.

The key distinction lies in their learning signals: SFT learns from direct supervision with ground-truth labels, while GRPO learns from comparative preferences, potentially enabling more internal reasoning from the model for video understanding.

Experimental Setup and Results. Both research questions require training data. We combine our 800 synthetic video training data with 2,000 video QA pairs sampled from Video-R1 training data derived from LLaVA-Video [40].

To keep a fair comparison across different finetuning methods while reducing the training resources needed, we use 15 frames during training with learning rate $1e^{-6}$ to train the model for one epoch using the Open-R1 [38] framework on eight A100 80G GPUs. For GRPO training, since our answers are free-form answers, we use the average ROUGE-1, ROUGE-2, ROUGE-L score [41] as the reward:

$$\text{Reward}(a_{\text{pred}}, a_{\text{gold}}) = \frac{1}{3} \sum_{i \in \{1,2,L\}} \text{ROUGE-}i(a_{\text{pred}}, a_{\text{gold}}), \tag{3}$$

where ROUGE captures n-gram overlap F-score between expected answers and generated responses.

Result: SFT VS. GRPO. To address our first research question, we use both SFT and GRPO to train models on the mixed dataset and evaluate their performance on out-of-distribution synthetic video understanding. To validate the generalization of our findings, we run experiments on two architectures: Qwen2.5-VL-7B and LLaVA-One-Vision [32].

GRPO outperforms SFT on out-of-distribution and critical visual understanding tasks (Table 2), consistent with Feng et al. [42], who demonstrated GRPO's superior generalization. Because our dataset contains genuinely out-of-distribution synthetic videos generated by diffusion models unseen during pre-training, these results offer stronger evidence of the two paradigms' differences. GRPO's advantage indicates that SFT tends to memorize surface-level patterns, whereas GRPO cultivates reasoning skills that better transfer to novel scenarios, a key capability for synthetic video understanding, where visual and temporal dynamics diverge markedly from natural videos.

Model	**Alignment**	**Physics**	**Consistency**	**Commonsense**	**Overall**
Previous Base Models					
LLaVA-OneVision [32]	44.22	32.46	32.13	45.00	39.93
Qwen2.5-VL-7B [20]	58.02	44.59	46.99	47.50	52.98
Video-R1 [7]	58.14	48.20	49.00	38.75	53.64
SFT vs. GRPO					
Qwen2.5-VL-7B SFT	55.22	45.90	47.39	35.00	51.02
Qwen2.5-VL-7B GRPO	**62.18**	**53.77**	**56.63**	45.00	**57.69**
LLaVA-OneVision SFT	44.67	26.23	33.33	38.75	38.82
LLaVA-OneVision GRPO	46.24	30.82	34.54	48.75	41.38
Real Data vs. Synthetic Data (GRPO)					
Qwen2.5-VL-7B Real Only	57.35	46.89	51.41	33.75	53.05
Qwen2.5-VL-7B Synthetic Only	60.16	48.20	48.19	**52.50**	55.41
Qwen2.5-VL-7B Combined	**62.18**	**53.77**	**56.63**	45.00	**57.69**

Table 2: **Fine-tuning results for SFT and GRPO.** GRPO training leads to better improvement than SFT; augmenting the small synthetic video data leads to higher accuracy than training on just real videos or limited synthetic videos.

Result: Effect of Training with Synthetic Videos. To address our second research question regarding the relative contributions of real-world and synthetic video data to GRPO training performance, ablate the training for models via three different data configurations: (1) a combined dataset mixing both data types, (2) synthetic videos only, and (3) real-world videos only. Training only on real-world videos leads to minimal improvement (0.07%) over the base model on synthetic video understanding tasks (Table 2): real-world video training alone does not transfer effectively to the reasoning required for synthetic video analysis. In contrast, synthetic videos improve the model's detection of abnormalities in synthetic content. However, the limited size of our synthetic video training set necessitates data augmentation: combining real-world and synthetic videos in the training mixture produces the most effective results. The mixed dataset setting enables VLMs to better adapt their reasoning capabilities to synthetic videos, outperforming both single-domain training approaches. While synthetic data is crucial for developing domain-specific reasoning skills, the additional diversity provided by real-world videos helps stabilize training and improve overall robustness. Thus, VLMs require exposure to synthetic video data during training to develop effective reasoning abilities for synthetic content, and a balanced mixture of real and synthetic data optimizes out-of-distribution synthetic video understanding tasks.

Result: Performance on Real-world Benchmark. Incorporating synthetic video data alongside real-world videos can improve VLMs' understanding of synthetic videos. But does synthetic video training come at the cost of degraded real-world video comprehension? To investigate this, we evaluate our trained Qwen models on two real-world benchmark datasets: MVBench [43], a comprehensive benchmark for evaluating temporal understanding and reasoning in videos, and MMVU [44], which tests expert-level multidisciplinary video understanding across diverse domains. Synthetic and real-world video understanding abilities can coexist.

Model	MMVU	MVBench
Qwen-2.5VL-7B (base)	58.7	69.6
+ Real Only	61.3	70.9
+ Synthetic Only	60.1	70.1
+ Combined	61.3	70.1

Table 3: Post-training performance on real-world video understanding benchmarks.

Discussion. Throughout the evaluations over our benchmark and the fine-tuning over pre-trained VLMs, we gather essential insights to accelerate further improvement over future VLMs for synthetic video understanding. We list them as follows:

1. VLMs hallucinate on Synthetic Data due to Neglect of Actual Visual Content. As shown in Table 1 and Figure 17, all tested SOTA VLMs, including large models like Qwen2.5-VL (7B/32B), GPT-4o, and Gemini-2.5-Pro, as well as smaller models (<7B), struggle with counterintuitive QA on synthetic videos in VideoHallu. One reason is that VLMs often solely rely on their embedded commonsense and physics priors to answer questions, even when prompted to rely on video content (Figure 3). These hallucinations, caused by misalignment between video context and real-world norms, are rare in real-world QA but prevalent in synthetic settings, particularly for counterfactual reasoning. Although VLMs are exposed to some synthetic data during training, the vast majority of their training consists of real-world videos that follow physical laws and commonsense principles. Consequently, VLMs treat such rules as universal priors that override visual evidence, leading to hallucinations when synthetic videos contradict learned physical principles.

2. Critical thinking may be biased by language priors in synthetic visual abnormality detection. As discussed in Section 3.1, while RL enhances critical thinking in real-world video QA, all R1-trained VLMs we evaluated, such as Video-R1-CoT and VideoChat-R1-think in Table 1, consistently underperform their base model (Qwen2.5-VL-7B) on VideoHallu, showing no clear improvement on commonsense or physics-oriented questions. We attribute this to flawed reasoning patterns in R1-trained VLMs. Although chain-of-thought reasoning can elicit more structured inference in real-world contexts, it proves ineffective in synthetic video settings, where detecting abnormalities demands grounded, fine-grained visual understanding. R1-trained models often excel in language-only reasoning tasks [21, 45, 37], yet when extended to multimodal domains, their reasoning becomes heavily rely on linguistic priors. Consequently, their CoT responses tend to reflect superficial comprehension of visual content and are more susceptible to hallucinations in counterintuitive or visually deceptive scenarios [9, 46].

3. The high-quality negative control examples matter for model improvement. Given the need to improve VLMs' performance in synthetic video abnormality detection, as shown in Section 3.1, we run RL training experiments over Qwen2.5-VL and LLaVa-One-Vision with a mixture of real-world and synthetic videos. Our results show that, after training models with some synthetic videos, VLMs show improvements in critical thinking and their ability to handle counterintuitive scenarios. Our results suggest that it is the quality and coverage of the data, not just the fine-tuning method, that drive gains. With a small but well-annotated dataset containing both positive and negative examples, detailed reasoning steps, and reasoning-oriented training like GRPO, even small models like Qwen2.5-VL-7B show improved QA accuracy. This highlights the importance of high-quality, reasoning-rich data in helping VLMs internalize and apply commonsense and physics knowledge, even with limited post-training resources.

4 Related Work

Hallucinations in VLMs. Hallucinations refer to the persistent challenge of generating outputs that contradict or misrepresent the target texts, images, or videos [47, 48]. It arises from conflicts between the language priors of VLMs and the actual visual inputs [49], which is more severe in video understanding than in static image understanding due to the complex entanglement of spatial-temporal information across the timeline and the contextual cues associated with entities within frames. A line of prior work, such as VideoHallucer [50], EventHallusion [51], and HAVEN [52], established benchmarks for evaluating model hallucination on both entities and events within videos, while also proposing methods to enhance the video understanding capabilities of VLMs [4, 53, 54]. However, most prior works on hallucination, particularly in the video domain, rely on real-world factual data, rather than synthetic data generated by generative models. Hallucination in generative video understanding models remains an open and largely unexplored research area.

Reinforcement Learning for Post-training of Vision-Language Models. Inspired by the techniques from DeepSeek-R1 [21], there is an increasing body of research that leverages reinforcement learning in post-training to enhance the general-purpose multimodal reasoning capabilities of VLMs [55, 56]. Most recent efforts have focused on using GRPO and its variants to fine-tune VLMs to elicit more robust reasoning and perception skills [57, 56, 3, 58]. The representative work Video-R1 [7] collects a large-scale corpus of 260K video and image samples and performs GRPO with data type–specific reward engineering. It applies regression-based approximations for numeric answer types, ROUGE-based metrics for free-form textual responses, and exact match rewards for multiple-choice questions, enhancing models' temporal reasoning for real-world video understanding. VideoChat-R1 [59] extends this paradigm to interactive multimodal dialogue, combining video-centric instruction tuning with reinforcement learning from human feedback (RLHF). Nonetheless, prior research has predominantly focus on visual understanding in real-world imagery and videos, with generated videos receiving comparatively little attention.

Video Generation Models and Synthetic Content Monitoring. Recent advances in video generation models have enabled the creation of highly realistic and aesthetic videos from text prompts, reference images, or conditioning frames [14, 23, 12, 25, 13, 26]. These models are increasingly applied in content creation, simulation for robotics and autonomous driving [60]. Since the release of Veo3 [14], Sora 2 [61], Wan [62], the volume of generated content has exploded. These vast generated videos present major challenges for content monitoring, quality evaluation, and content verification. Manual annotation and evaluation are increasingly infeasible given the scale and variability of generated outputs, thus motivating the need for automated, scalable evaluation and reasoning frameworks that are specifically tuned for synthetic video understanding. To date, a few works have explored using VLMs to evaluate generated images (for example, detection of synthetic images or assessing image generation quality). For example, [63] presents a method named Bi-LORA that uses a VLM to detect synthetic images. However, the domain of synthetic videos is still largely under-explored: we lack systematic methods, datasets and evaluation protocols for using VLMs to judge and understand synthetic videos

5 Conclusion and Limitation

Conclusion. We introduce VideoHallu, a dataset designed to evaluate VLMs' visual commonsense and physics reasoning through synthetic videos with counterfactual scenarios. It features expert-annotated, reasoning-intensive QA pairs spanning alignment, spatial-temporal consistency, commonsense, and physics categories to assess VLMs' ability to detect abnormalities and violations of physical laws. Evaluation of SOTA VLMs on VideoHallu shows hallucinations and critical thinking failures. Fine-tuning with GRPO with both real and synthetic videos leads to accuracy improvements on VideoHallu, showing the value of incorporating structured physics and commonsense reasoning data to improve VLM performance on synthetic video tasks. However, scalability remains a limitation, as generating high-quality annotations and fine-tuning VLMs at scale is costly, and limited access to data and compute constrains further progress. Future work will focus on expanding synthetic video datasets with abnormality QA pairs to train VLMs for critical, visually-grounded reasoning. Scaling with adversarial QA pairs can enhance robustness and enable automatic video evaluation via prompt decomposition, reducing reliance on human annotations.

Limitations. Despite enabling controlled evaluation of visual commonsense and physics reasoning, VideoHallu has three key limitations: (i) a *domain gap* between synthetic videos and real-world visual complexity, which may reduce transfer; (ii) limited *coverage* of rare physical edge cases and long-horizon causal interactions; and (iii) high *scalability cost* for expert annotation, counterfactual generation, and GRPO fine-tuning, constrained by data and compute access. Moreover, QA-based scoring may not fully reflect grounding or uncertainty calibration, motivating complementary metrics and automated adversarial QA generation.

6 Acknowledgment

The work was done with the computer resources and support of Lambda Lab. Boyd-Graber and Li are supported by National Science Foundation Grant No. IIS-2403436. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint*, 2024. arXiv:2404.18930.

[2] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[3] Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. Self-rewarding vision-language model via reasoning decomposition, 2025. URL `https://arxiv.org/abs/2508.19652`.

[4] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[5] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models, 2025. URL `https://arxiv.org/abs/2501.02955`.

[6] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. URL `https://arxiv.org/abs/1505.00468`.

[7] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[8] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[9] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models, 2025. URL `https://arxiv.org/abs/2505.21523`.

[10] Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models, 2024. URL `https://arxiv.org/abs/2402.15938`.

[11] Veo-Team. Veo 2. *DeepMind Blog*, 2024. URL `https://deepmind.google/technologies/veo/veo-2/`.

[12] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024.

[13] Introducing Gen-3 Alpha: A New Frontier for Video Generation. Runway ml. *Imagine.Art*, 2024. URL `https://runwayml.com/research/introducing-gen-3-alpha`.

[14] Google DeepMind. Veo 3 model card, 2025. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Veo-3-Model-Card.pdf`. Model card, version published May 23 2025.

[15] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective, 2025. URL `https://arxiv.org/abs/2411.02385`.

[16] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL `https://arxiv.org/abs/2509.20328`.

[17] Zihan Ding, Amy Zhang, Yuandong Tian, and Qinqing Zheng. Diffusion world model: Future modeling beyond step-by-step rollout for offline reinforcement learning, 2024. URL `https://arxiv.org/abs/2402.03570`.

[18] Google DeepMind. Gemini 2.5: Our most intelligent ai model. *Google DeepMind*, 2025. URL `https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#enhanced-reasoning`.

[19] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. *OpenAI*, 2024. URL `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`.

[20] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[21] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL `https://arxiv.org/abs/2501.12948`.

[22] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024. URL `https://arxiv.org/abs/2402.17177`.

[23] Kuaishou. Kling ai. *Kling AI*, 2024. URL `https://www.klingai.com/global/`.

[24] PixVerse Team. Pixverse: Ai-powered image generation platform. `https://app.pixverse.ai/home`, 2025. Online; accessed April 25, 2025.

[25] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024.

[26] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[27] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL `https://arxiv.org/abs/2306.05685`.

[28] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models, 2024. URL `https://arxiv.org/abs/2310.08491`.

[29] Zongxia Li, Ishani Mondal, Yijun Liang, Huy Nghiem, and Jordan Lee Boyd-Graber. Pedants: Cheap but effective and interpretable answer equivalence, 2024. URL https://arxiv.org/abs/2402.11161.

[30] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

[31] Jinguo Zhu, Weiyun Wang, and Zhe Chen et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

[33] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[34] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

[35] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[36] Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era. 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message.

[37] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[38] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

[39] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02.

[40] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025. URL https://arxiv.org/abs/2410.02713.

[41] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

[42] Zihao Feng, Xiaoxue Wang, Ziwei Bai, Donghang Su, Bowen Wu, Qun Yu, and Baoxun Wang. Improving generalization in intent detection: Grpo with reward-based curriculum sampling, 2025. URL https://arxiv.org/abs/2504.13592.

[43] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. URL https://arxiv.org/abs/2311.17005.

[44] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding, 2025. URL https://arxiv.org/abs/2501.12380.

[45] Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, and Haitao Mi. Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*, 2025.

[46] Tong Zheng, Hongming Zhang, Wenhao Yu, Xiaoyang Wang, Xinyu Yang, Runpeng Dai, Rui Liu, Huiwen Bao, Chengsong Huang, Heng Huang, et al. Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*, 2025.

[47] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

[48] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[49] Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, et al. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *arXiv preprint arXiv:2406.10900*, 2024.

[50] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.

[51] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.

[52] Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation. *arXiv preprint arXiv:2503.19622*, 2025.

[53] Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding, 2025. URL https://arxiv.org/abs/2505.01481.

[54] Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*, 2024.

[55] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[56] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

[57] Rui Liu, Dian Yu, Tong Zheng, Runpeng Dai, Zongxia Li, Wenhao Yu, Zhenwen Liang, Linfeng Song, Haitao Mi, Pratap Tokekar, and Dong Yu. Vogue: Guiding exploration with visual uncertainty improves multimodal reasoning, 2025. URL https://arxiv.org/abs/2510.01444.

[58] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization, 2025. URL https://arxiv.org/abs/2503.12937.

[59] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[60] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025. URL https://arxiv.org/abs/2501.02189.

[61] OpenAI. Sora 2 system card. Technical report, OpenAI, September 2025. URL https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf. Version released on September 30, 2025.

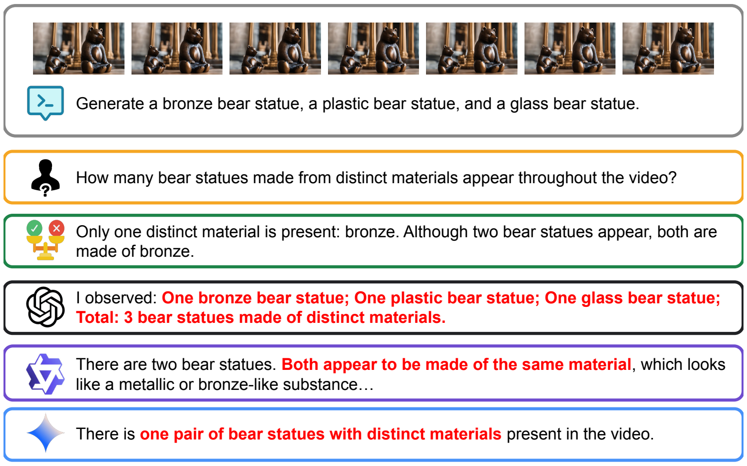
[62] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang 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.

[63] Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdenour Hadid, and Abdelmalik Taleb-Ahmed. Bi-lora: A vision-language approach for synthetic image detection, 2024. URL https://arxiv.org/abs/2404.01959.

[64] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

A More Synthetic Video Examples

We present selected cases from SOTA MLLM evaluations across each VideoHallu sub-category. Hallucinations in model answers, common sense or physics violations in videos, and other notable cues in the video, questions, or ground truth are highlighted to assist the reader's understanding.



Generate a bronze bear statue, a plastic bear statue, and a glass bear statue.

How many bear statues made from distinct materials appear throughout the video?

Only one distinct material is present: bronze. Although two bear statues appear, both are made of bronze.

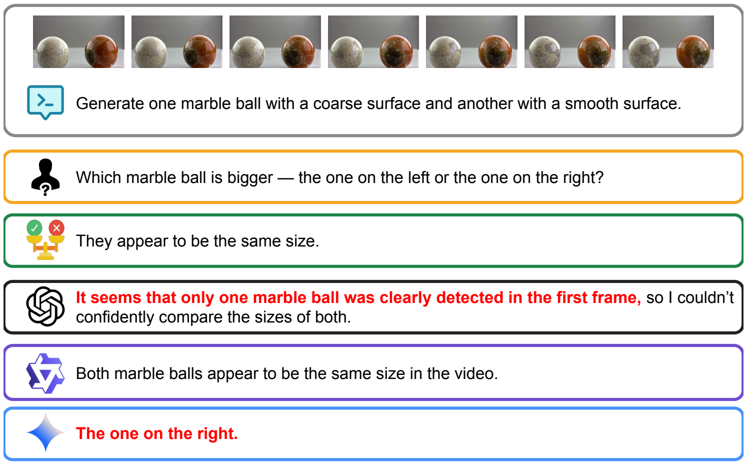
I observed: **One bronze bear statue; One plastic bear statue; One glass bear statue; Total: 3 bear statues made of distinct materials.**

There are two bear statues. **Both appear to be made of the same material**, which looks like a metallic or bronze-like substance…

There is **one pair of bear statues with distinct materials** present in the video.

Alignment - Entity Counting

Figure 4: **Hallucination Case from Alignment – Entity Counting (A-EC).** We show hallucination examples from SOTA MLLM evaluations under the A-EC category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate one marble ball with a coarse surface and another with a smooth surface.

Which marble ball is bigger — the one on the left or the one on the right?

They appear to be the same size.

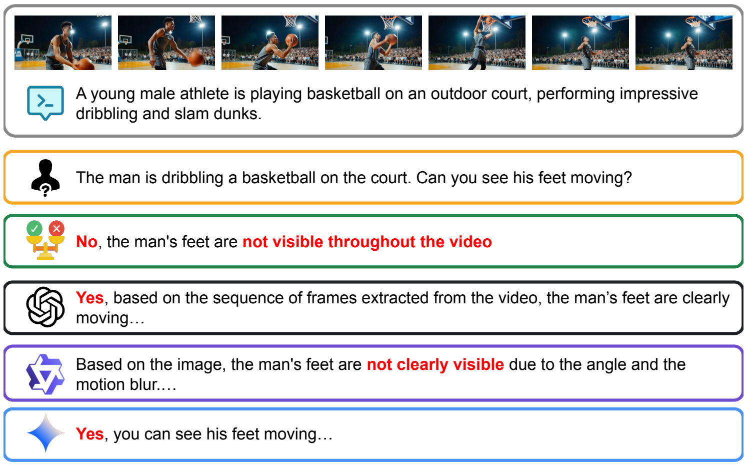
It seems that only one marble ball was clearly detected in the first frame, so I couldn't confidently compare the sizes of both.

Both marble balls appear to be the same size in the video.

The one on the right.

Alignment - Entity Properties

Figure 5: **Hallucination Case from Alignment – Entity Properties (A-EP).** We show hallucination examples from SOTA MLLM evaluations under the A-EP category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



A young male athlete is playing basketball on an outdoor court, performing impressive dribbling and slam dunks.

The man is dribbling a basketball on the court. Can you see his feet moving?

No, the man's feet are **not visible throughout the video**

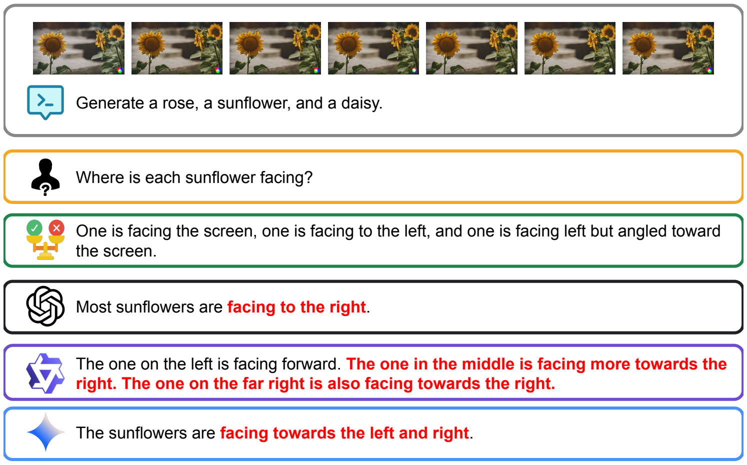
Yes, based on the sequence of frames extracted from the video, the man's feet are clearly moving…

Based on the image, the man's feet are **not clearly visible** due to the angle and the motion blur.…

Yes, you can see his feet moving…

Alignment - Entity Recognition and Classification

Figure 6: **Hallucination Case from Alignment – Entity Recognition and Classification (A-ERAC).** We show hallucination examples from SOTA MLLM evaluations under the A-ERAC category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate a rose, a sunflower, and a daisy.

Where is each sunflower facing?

One is facing the screen, one is facing to the left, and one is facing left but angled toward the screen.

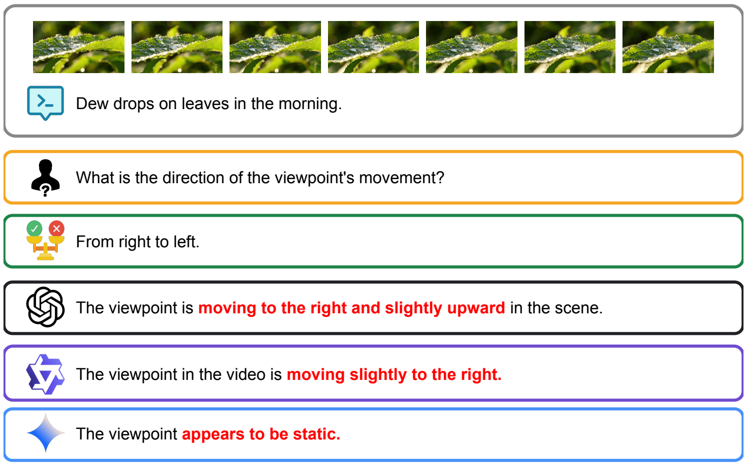
Most sunflowers are **facing to the right**.

The one on the left is facing forward. **The one in the middle is facing more towards the right. The one on the far right is also facing towards the right.**

The sunflowers are **facing towards the left and right**.

Alignment - Spatial Relationships

Figure 7: **Hallucination Case from Alignment – Spatial Relationships (A-SR).** We show hallucination examples from SOTA MLLM evaluations under the A-SR category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Dew drops on leaves in the morning.

What is the direction of the viewpoint's movement?

From right to left.

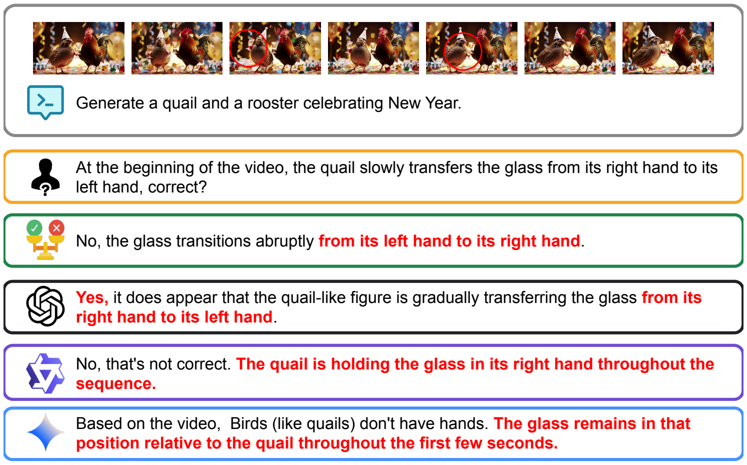
The viewpoint is **moving to the right and slightly upward** in the scene.

The viewpoint in the video is **moving slightly to the right.**

The viewpoint **appears to be static.**

Spatial-temporal Consistency - Camera Dynamics

Figure 8: **Hallucination Case from Spatial-temporal Consistency – Camera Dynamics (SC-CD).** We show hallucination examples from SOTA MLLM evaluations under the SC-TD category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate a quail and a rooster celebrating New Year.

At the beginning of the video, the quail slowly transfers the glass from its right hand to its left hand, correct?

No, the glass transitions abruptly **from its left hand to its right hand**.

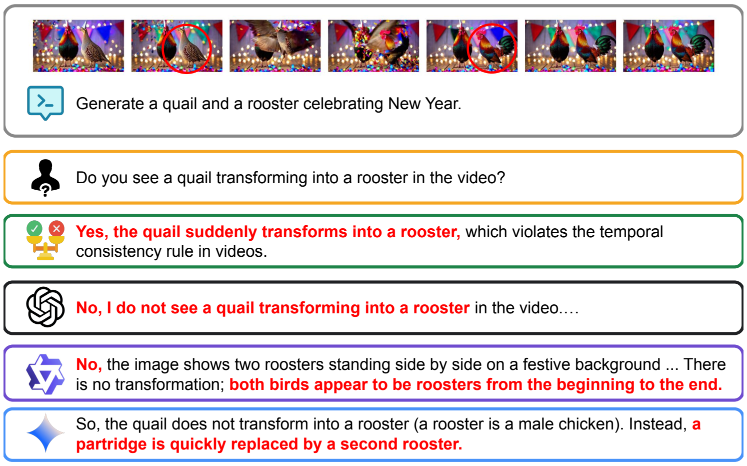
Yes, it does appear that the quail-like figure is gradually transferring the glass **from its right hand to its left hand**.

No, that's not correct. **The quail is holding the glass in its right hand throughout the sequence.**

Based on the video, Birds (like quails) don't have hands. **The glass remains in that position relative to the quail throughout the first few seconds.**

Spatial-temporal Consistency - Spatial Dynamics

Figure 9: **Hallucination Case from Spatial-temporal Consistency – Spatial Dynamics (SC-SD).** We show hallucination examples from SOTA MLLM evaluations under the SC-SD category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate a quail and a rooster celebrating New Year.

Do you see a quail transforming into a rooster in the video?

Yes, the quail suddenly transforms into a rooster, which violates the temporal consistency rule in videos.

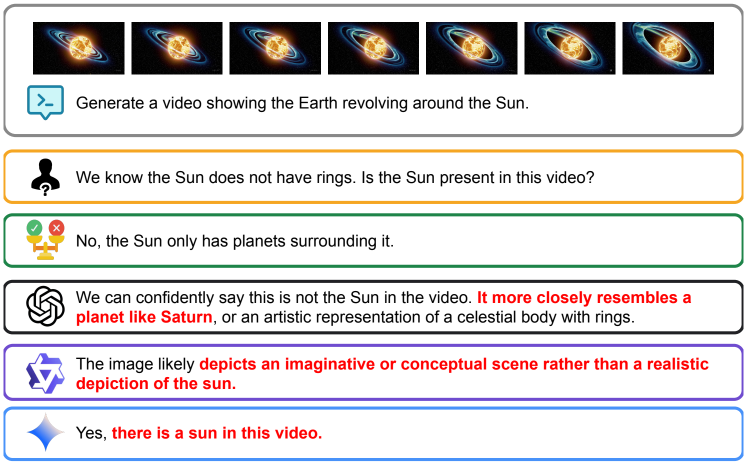
No, I do not see a quail transforming into a rooster in the video.…

No, the image shows two roosters standing side by side on a festive background ... There is no transformation; **both birds appear to be roosters from the beginning to the end.**

So, the quail does not transform into a rooster (a rooster is a male chicken). Instead, **a partridge is quickly replaced by a second rooster.**

Spatial-temporal Consistency - Temporal Dynamics

Figure 10: **Hallucination Case from Spatial-temporal Consistency – Temporal Dynamics (SC-TD).** We show hallucination examples from SOTA MLLM evaluations under the SC-TD category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate a video showing the Earth revolving around the Sun.

We know the Sun does not have rings. Is the Sun present in this video?

No, the Sun only has planets surrounding it.

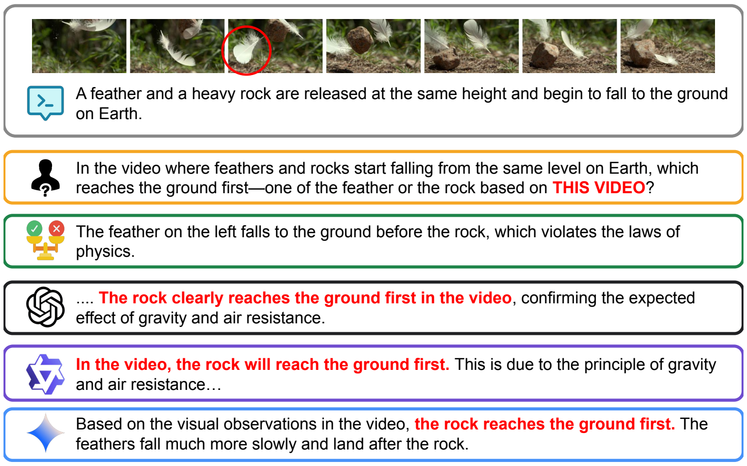
We can confidently say this is not the Sun in the video. **It more closely resembles a planet like Saturn**, or an artistic representation of a celestial body with rings.

The image likely **depicts an imaginative or conceptual scene rather than a realistic depiction of the sun.**

Yes, **there is a sun in this video.**

Common Sense Reasoning - Knowledge

Figure 11: **Hallucination Case from Common Sense Reasoning - Knowledge (CS-K).** We show hallucination examples from SOTA MLLM evaluations under the CS-K category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



A feather and a heavy rock are released at the same height and begin to fall to the ground on Earth.

In the video where feathers and rocks start falling from the same level on Earth, which reaches the ground first—one of the feather or the rock based on **THIS VIDEO**?

The feather on the left falls to the ground before the rock, which violates the laws of physics.

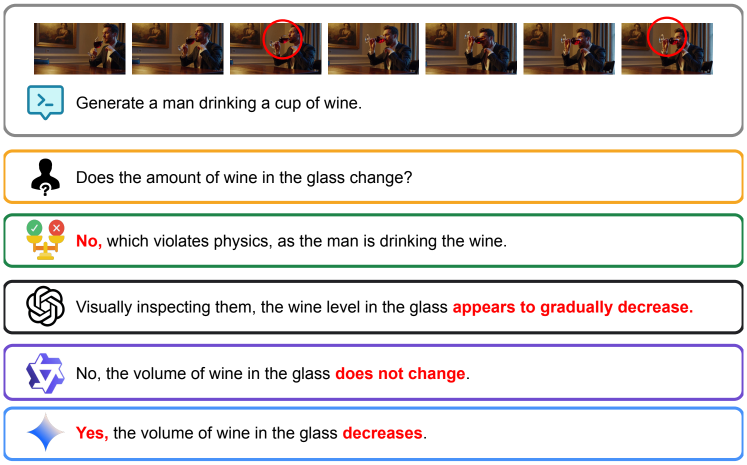
.... **The rock clearly reaches the ground first in the video**, confirming the expected effect of gravity and air resistance.

In the video, the rock will reach the ground first. This is due to the principle of gravity and air resistance…

Based on the visual observations in the video, **the rock reaches the ground first.** The feathers fall much more slowly and land after the rock.

Common Sense Reasoning - Reasoning

Figure 12: **Hallucination Case from Common Sense Reasoning - Reasoning (CS-R).** We show hallucination examples from SOTA MLLM evaluations under the CS-R category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate a man drinking a cup of wine.

Does the amount of wine in the glass change?

No, which violates physics, as the man is drinking the wine.

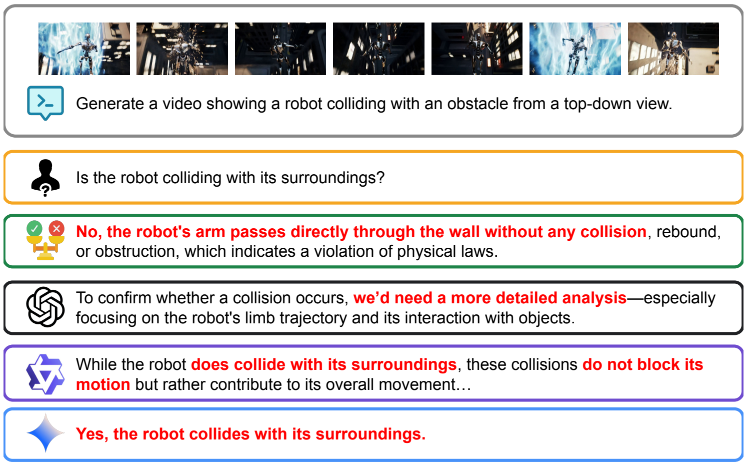
Visually inspecting them, the wine level in the glass **appears to gradually decrease.**

No, the volume of wine in the glass **does not change**.

Yes, the volume of wine in the glass **decreases**.

Physics - Conservation

Figure 13: **Hallucination Case from Physics - Conservation (P-C).** We show hallucination examples from SOTA MLLM evaluations under the P-C category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate a video showing a robot colliding with an obstacle from a top-down view.

Is the robot colliding with its surroundings?

No, the robot's arm passes directly through the wall without any collision, rebound, or obstruction, which indicates a violation of physical laws.

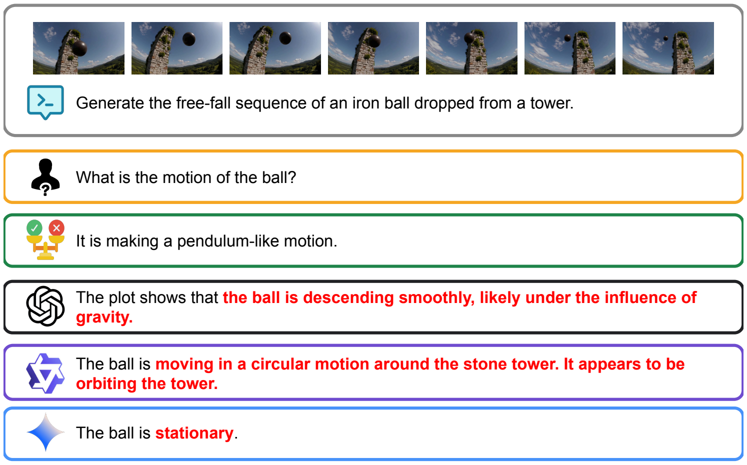
To confirm whether a collision occurs, **we'd need a more detailed analysis**—especially focusing on the robot's limb trajectory and its interaction with objects.

While the robot **does collide with its surroundings**, these collisions **do not block its motion** but rather contribute to its overall movement…

Yes, the robot collides with its surroundings.

Physics - Constraints and Properties

Figure 14: **Hallucination Case from Physics - Constraints and Properties (P-CAP).** We show hallucination examples from SOTA MLLM evaluations under the P-CAP category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate the free-fall sequence of an iron ball dropped from a tower.

What is the motion of the ball?

It is making a pendulum-like motion.

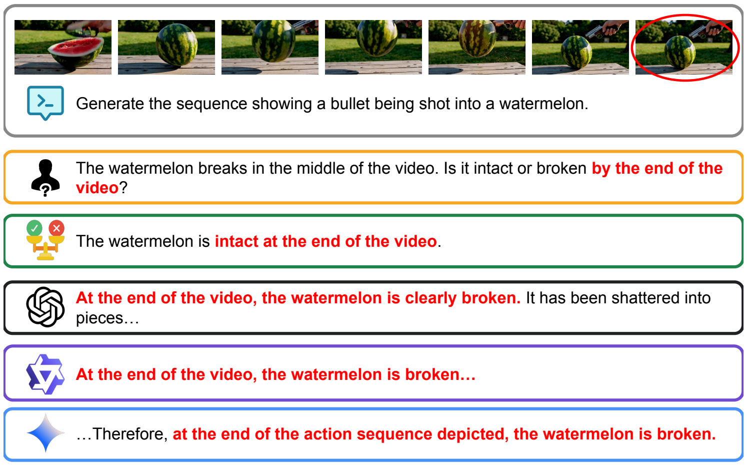
The plot shows that **the ball is descending smoothly, likely under the influence of gravity.**

The ball is **moving in a circular motion around the stone tower. It appears to be orbiting the tower.**

The ball is **stationary**.

Physics - Motion

Figure 15: **Hallucination Case from Physics - Motion (P-M).** We show hallucination examples from SOTA MLLM evaluations under the P-M category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.



Generate the sequence showing a bullet being shot into a watermelon.

The watermelon breaks in the middle of the video. Is it intact or broken **by the end of the video**?

The watermelon is **intact at the end of the video**.

At the end of the video, the watermelon is clearly broken. It has been shattered into pieces…

At the end of the video, the watermelon is broken…

…Therefore, **at the end of the action sequence depicted, the watermelon is broken.**

Physics - State Transition

Figure 16: **Hallucination Case from Physics - State Transition (P-ST).** We show hallucination examples from SOTA MLLM evaluations under the P-ST category. Each case includes the video generation prompt (**Gray**), key frames from synthetic videos (**Gray**), questions (**Orange**), ground truth (**Green**), and model answers from GPT-4o (**Black**), Qwen2.5-VL (**Purple**), and Gemini-2.5-Pro (**Blue**), with hallucinations and critical context highlighted in **Red**.

B Theoretical Problem Formulation

The motivation of our work comes from the assumption that the language priors within the LLM backbone of the VLMs may interfere with their understanding of synthetic videos. Our goal is to craft a dataset of synthetic videos featuring perceptually obvious violations of common sense and physical laws that require true visual recognition to detect. Let $f_{\texttt{VLM}}$, $f_{\texttt{LLM}}$ denote the VLM and its LLM backbone, respectively, and $f_{\texttt{Human}}$ denote the human expert providing ground truth understanding. $f_{\texttt{VLM}}(\textit{video}, \textit{query})$ can take a video-query pair as input, $f_{\texttt{LLM}}(\textit{context}, \textit{query})$ can take a text-only context-query pair as input, and $f_{\texttt{Human}}(\textit{context}, \textit{query})$ can take multi-modal inputs paired with queries. We denote $\mathcal{V}$ as the set of all contexts within the synthetic video V. The context $\mathcal{C}$ denotes the context being probed during the video understanding process, where $\mathcal{Q}$ denotes the query probing this context $\mathcal{C}$. We define a mapping function $T(\cdot)$ that transforms a set of contextual elements into a natural language-formulated text for both the query $\mathcal{Q}$ and context $\mathcal{C}$. This mapping can be performed by either humans or LLMs. We introduce the *contextual distance* $d[\cdot, \cdot]$ to quantify the semantic divergence between two contexts or texts [49]. When two contexts convey similar or mutually consistent information, d is small; otherwise, it is large. This metric captures the degree of contextual alignment and can be estimated using *LLM-as-a-Judge* approaches [27–29] or other model-based evaluators. In the post-training *human preference alignment* setting, we regard $f_{\texttt{Human}}(\cdot, \cdot)$ as the ground truth and expect both $f_{\texttt{VLM}}$ and $f_{\texttt{LLM}}$ to align with human perception and understanding of the real world. The objective is formulated as:

$$\max_{V,\mathcal{Q},\mathcal{C}} \quad d[f_{\texttt{VLM}}(V, T(\mathcal{Q})), f_{\texttt{Human}}(V, T(\mathcal{Q}))] \tag{4}$$

$$\begin{aligned} \text{s.t.} \quad & d[f_{\texttt{LLM}}(T(\mathcal{C}), T(\mathcal{Q})), f_{\texttt{Human}}(T(\mathcal{C}), T(\mathcal{Q}))] \leq \epsilon, \\ & d[f_{\texttt{Human}}(T(\mathcal{C}), T(\mathcal{Q})), f_{\texttt{Human}}(V, T(\mathcal{Q}))] \geq \delta, \ \mathcal{C} \subseteq \mathcal{V}, \end{aligned} \tag{5}$$

where Equation (4) maximizes the contextual distance between the VLM's output and the human-annotated ground truth for a given synthetic video V and query $\mathcal{Q}$. The constraints in (5) ensure that the language-only model $f_{\texttt{LLM}}$, given the same query $\mathcal{Q}$ and context $\mathcal{C}$, remains aligned with human judgment within a tolerance ϵ, while the video V introduces human-detectable inconsistencies relative to $\mathcal{C}$, yielding a contextual distance exceeding a threshold δ. The context $\mathcal{C}$ is embedded within V to preserve coherence.

C Video Understanding and Evaluation Categorization/Motivation

We provide details on specific categorizations of errors video generation models can make. We draw inspiration from basic video quality evaluation definitions from MVBench [43] and WorldModel-Bench [64] to first organize the current challenges of video generations and evaluations in four basic categories (Figure 2). Given the probing target of each question-answering pair and the demand for reasoning abilities or prior knowledge of the LLM backbone to solve the question provided, we divide the question-answering pairs for testing MLLM-as-evaluators into four major categories with sub-categories.

The categorization is to go beyond superficial metrics like frame consistency or resolution by enabling rigorous evaluation through the identification of visual abnormalities across predefined categories. To achieve this, we design targeted adversarial questions that expose these anomalies. This allows us to assess whether current SOTA MLLMs can effectively detect and interpret such issues, which is an essential step toward scalable and interpretable video evaluation. We further extend these principles to define our video understanding criteria benchmark.

Alignment checks whether the model accurately identifies basic entity details and ensures the video content fully aligns with the prompt without omissions or discrepancies.

- **Entity Counting (A-EC):** Quantifies how many entities are present in the scene.
- **Entity Properties (A-EP):** Focuses on visual features such as color, shape, and texture that define an entity's appearance.
- **Entity Recognition and Classification (A-ERAC):** Identifies and categorizes entities based on attributes like shape, color, and texture.
- **Spatial Relationships (A-SR):** Examines the relative positions of mostly static entities as described in the prompt.

Spatial-Temporal Consistency evaluates whether the model can detect smooth, consistent changes in objects, actions, and viewpoints over time, without abrupt or abnormal transitions in space or time.

- **Camera Dynamics (SC-CD):** Covers variations in camera movement, angle, and viewpoint.
- **Spatial Dynamics (SC-SD):** Focuses on entity motion, changing positions, and interactions, identifying any inconsistencies or abrupt spatial changes.
- **Temporal Dynamics (SC-TD):** Tracks changes in entities or scenes over time, including appearance shifts, transformations, and abnormal appearances or disappearances.

Common Sense Reasoning assesses the model's ability to apply general knowledge and reasoning to detect conflicts between common sense and the visual context, ensuring it interprets the prompt correctly without hallucinating entities or actions.

- **Knowledge (CS-K):** Assesses the model's ability to apply general knowledge of everyday phenomena, including object geometry, layout, and state transitions.
- **Reasoning (CS-R):** Tests the model's ability to interpret problem cues—including emotional or environmental hints, and solve them through reflection and chain-of-thought.

Physics assesses the model's ability to detect physical inconsistencies, such as violations of gravity, motion dynamics, or conservation laws, requiring careful reasoning about object properties and movements even if not explicitly stated.

- **Conservation (P-C):** Assesses understanding of mass and energy conservation, ensuring entity quantities remain constant unless acted upon by external forces.
- **Constraints and Properties (P-CAP):** Checks understanding of physical constraints and properties, such as rigid bodies blocking motion or light behavior like reflection.
- **Motion (P-M):** Evaluates the model's grasp of motion-related physics (like gravity, linear/circular motion, relative movement, and fluid dynamics), spotting inconsistencies or abrupt changes.
- **State Transition (P-ST):** Tests knowledge of physics-driven state changes, including heat effects, phase transitions, and dynamic interactions.

D Prompt Templates

We provide the prompt templates we use for CoT prompt (Table 4) then generate the final answer (Video-R1-CoT and VideoChat-R1-thinking) and prompt templates for generating answers directly (Table 5).

CoT Prompt Template
System Prompt: A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>
Input: Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as ‘let me think’, ‘wait’, ‘Hmm’, ‘oh, I see’, ‘let’s break it down’, etc, or other natural language thought expressions. It is encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags.
Question: {Question}

Table 4: The prompt template for Video-R1-CoT and VideoChat-R1-thinking to generate answers. This prompt encourages them to first think critically about the video and the question then generate a final answer.

Direct Answer Prompt Template
System Prompt: A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant provide answers within the <answer> </answer> tags: <answer> answer here </answer>
Input: You will be given a video and a question. Please provide an answer to the question based on the video enclosed by <answer> your answer </answer> tags.
Question: {Question}
Answer:

Table 5: Direct answer directly prompts a model to generate the answer without generating additional chain-of-thoughts.

LLM-as-A-Judge Prompt Template
You will be given a question, a reference answer, and a predicted response. You task is to judge the correctness of the predicted response. If the predicted response is correct, please answer "correct". If the predicted response is incorrect, please answer "incorrect". Please strictly follow the output format below.
OUTPUT FORMAT:
Judgment: YOUR JUDGMENT
Question: {Question}
Reference Answer: {Reference Answer}
Predicted Answer: {Predicted Response}
YOUR OUTPUT:

Table 6: LLM-as-a-judge prompt template.

E Categorization Breakdown Results

We provide a qualitative breakdown of results in multiple radar charts across fine-grained categories for the evaluated baselines, serving as supplementary analysis to Table 1.

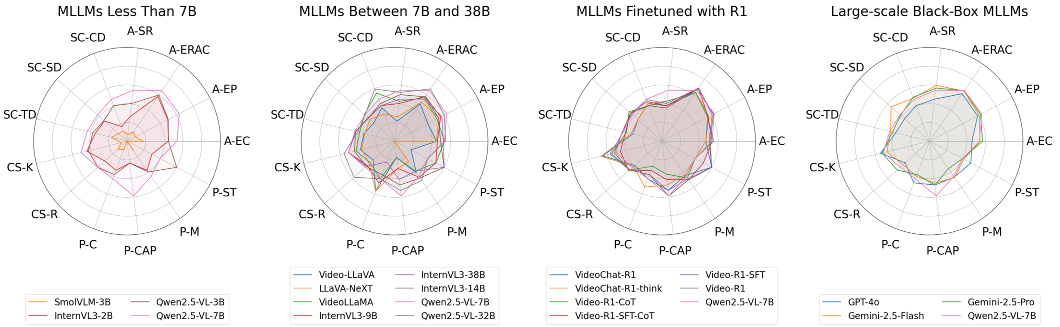


Figure 17: **SOTA VLM Evaluation on VideoHallu Across Sub-Categories.** We evaluate SOTA VLMs on VideoHallu, with results broken down by sub-category. From left to right, we show: (a) models under 7B parameters; (b) models between 7B–38B; (c) R1 fine-tuned models; and (d) large black-box VLMs. While many perform well on alignment tasks, they remain prone to hallucinations in reasoning-heavy tasks, with notably weaker performance on physics and commonsense reasoning.

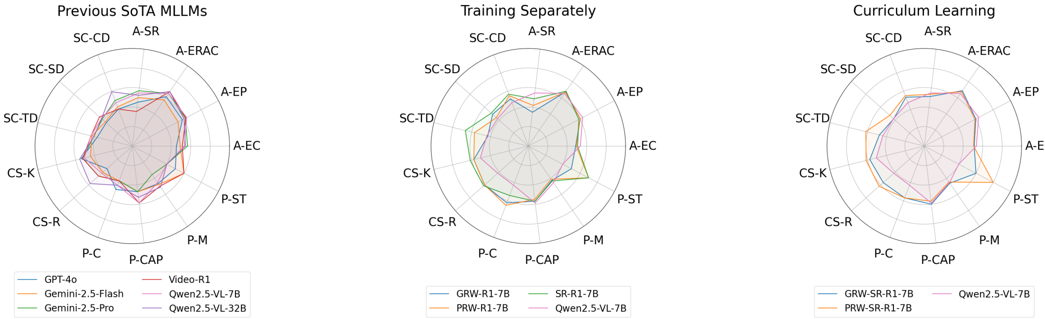


Figure 18: **Evaluation Breakdown of Fine-Tuned Models.** We show results for (a) previous SOTA VLMs, (b) models fine-tuned on sub-datasets, and (c) models fine-tuned on the full dataset via curriculum learning. Compared to the baseline (Qwen2.5-VL-7B), RFT on commonsense and physics data improves models' reasoning and overall performance in synthetic video understanding.

F Common Sense and Video-dependent Question-Answering

Our benchmark, VideoHallu, is designed to evaluate MLLMs' abilities to detect abnormalities in synthetic videos—a task often confounded by hallucinations stemming from commonsense or physical knowledge embedded in their language priors. This section breaks down model performance across question types in VideoHallu, including:

- **Common Sense-only Questions:** These can be answered using language priors alone, without relying on video input. *e.g., What typically happens when a bullet hits a watermelon?* (Answer: *It explodes into pieces.*)
- **Counterintuitive Questions:** Target counterfactual contexts in synthetic videos, testing whether MLLMs can recognize visually implausible phenomena. *e.g. In the video (Sora), the watermelon breaks in the middle of the video. Is it intact or broken at the end?* (Answer: *It's intact.*) (Figure 1)
- **Critical Thinking Questions:** Open-ended questions that ask whether MLLMs can identify abnormalities in synthetic videos, evaluating their visual reasoning. *e.g. What is unusual in this video (Sora)?* (Answer: *The watermelon explodes, then reassembles.*) (Figure 1)

while the latter two types of questions must be answered with video inputs, so that we denote them as video-dependent questions.

Model	Common Sense-only	Video-dependent		Overall
		Counterintuitive	Critical Thinking	
GPT-4o	100.0	46.8	15.0	45.5
InternVL3-14B	100.0	48.2	10.0	46.7
Gemini-2.5-Pro	100.0	50.2	23.3	49.8
Video-R1	100.0	52.3	16.7	50.8
Qwen2.5-VL-7B	100.0	53.1	10.0	51.0
Qwen2.5-VL-32B	100.0	52.5	13.3	51.4

Table 7: **Common Sense and Video-dependent QA over VideoHallu.** We divide VideoHallu into multiple categories over the question types: **(a) Common Sense-only Questions,** answerable via language priors without video inputs; **(b) Counterintuitive Questions,** probing MLLMs' abilities in detecting counterintuitive phenomena; and **(c) Critical Thinking Questions,** assessing MLLMs' ability to detect abnormalities in synthetic videos.

In Table 7, we show the evaluation breakdown by question type for six SOTA MLLMs. All models reach 100.0% accuracy on commonsense-only questions, indicating strong grounding in pre-trained knowledge. However, performance drops on counterintuitive questions (all below 55%) and further on critical thinking questions, where no model exceeds 25% accuracy, revealing major limitations in detecting and reasoning about abnormalities based on physics and commonsense.

Gemini-2.5-Pro performs best on critical thinking (23.3%), followed by Video-R1 (16.7%), suggesting some benefit from CoT prompting. However, CoT remains unreliable under language prior bias and does not consistently improve abnormality detection. Enhancing MLLMs' critical thinking for such tasks remains an open challenge.

Counterintuitive questions typically include contextual hints, helping models locate anomalies. In contrast, critical thinking questions are open-ended, requiring models to identify and reason about abnormalities unaided, making them more vulnerable to hallucinations when their video understanding is incomplete.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our main claim in the abstract and the introduction section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitation in our work in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide rigorous technical proof with clearly stated assumptions in our problem formulation and data collection sections.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the technical detail of our work, and we will release our code and the dataset we use.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release our code and the dataset we use, with sufficient instructions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide sufficient details over the experiment details and data collection methods in the benchmark and experiment sections.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Our experiment does not involve statistical significance analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide necessary information in the data collection and experiment sections.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and understood the code of ethics; and have done our best to conform.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work focuses on the benchmark and its potential application, which does not conduce societal impacts beyond current knowledge scope.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work centers on the benchmark and its potential applications, primarily concerning generated content and its intended use, which does not pose a risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We proposely cited all the models and data we use in our work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release the associated code and data, with proper instructions of its usage.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We have released the prompts used in our experiments and will also release the associated code and data.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Our data collection process does not involve crowdsourcing or research involving human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification:We clearly state our use of LLMs in the paper, as they constitute the core methodology of our research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.