Research Article

VideoHallu: Evaluating and Mitigating Multi-modal Hallucinations for Synthetic Videos

Zongxia Li[1], Xiyang Wu[1], Yubin Qin[1], Guangyao Shi[2], Hongyang Du[1], Dinesh Manocha[1], Tianyi Zhou[1], Jordan Lee Boyd-Graber[1]

1. University of Maryland, College Park, United States; 2. University of Southern California, United States

Synthetic video generation using foundation models has gained significant attention due to its realism and broad applications. However, while these models excel at generating visually coherent and high-quality video frames, they often overlook commonsense reasoning and physical law violations, leading to abnormal content.

Existing score-based evaluations like VideoScore[1] mainly focus on general video quality and do not take these abnormalities into account, and offer no explanations of the evaluation results. A more promising evaluation approach is to leverage multi-modal large language models (MLLMs) as interpretable video evaluators, following the approach of FactScore[2]. However, how well MLLMs can detect these abnormalities in synthetic videos is underexplored. Motivated by a more interpretable video generation evaluation, we introduce VideoHallu, a benchmark built from synthetic videos produced by popular models like Veo2[3], Sora[4], and Kling[5], paired with expert-crafted question-answering pair examples easily solvable with human-level perception and reasoning across multiple categories. We evaluate several State-of-the-Art (SoTA) MLLMs with our benchmark, including GPT-4o[6], Gemini-2.5-Pro[7], Qwen-2.5-VL[8], and forefront models like Video-R1[9] and VideoChat-R1[10]. Despite the strong performance of R1 MLLMs on real-world video benchmarks like MVBench[11] and MovieChat[12], these models still struggle and hallucinate on basic commonsense and physics reasoning tasks in synthetic videos, highlighting synthetic video hallucination as an underexplored challenge. Moreover, we post-train current SoTA MLLMs with Group Relative Policy Optimization (GRPO)[13] using both real-world and synthetic commonsense/physics datasets. Our results show improved overall accuracy compared to the base model, achieving the highest performance among all models, highlighting the importance of integrating high-quality counterexamples to enhance

 doi.org/10.32388/BXC6X1

commonsense and physics reasoning in MLLMs' language priors. Our data is available at https://github.com/zli12321/VideoHallu.git.

Zongxia Li and Xiyang Wu equally contributed to this work.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

1. Introduction

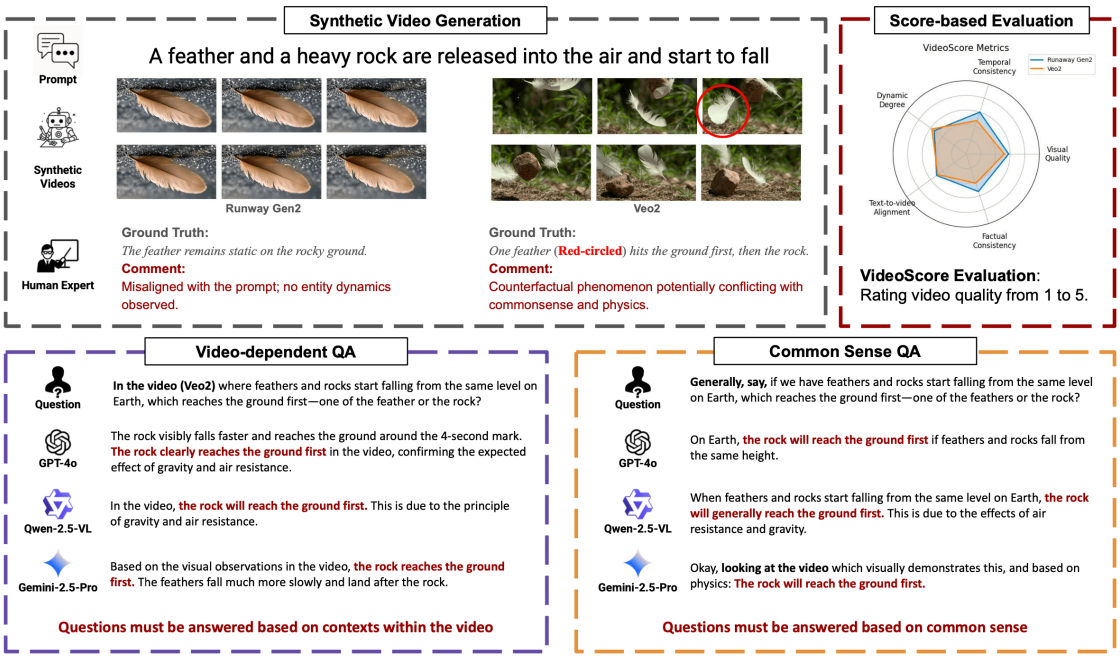


Figure 1. Hallucinations in Synthetic Video Understanding. Synthetic videos often exhibit counterfactual or common-sense-violating behaviors due to current video generation model limitations. While prior evaluation works mainly score video quality and consistency, our work, VideoHallu, probes hallucinations caused by misapplied common sense in video-dependent questions. We design both common sense QA (answerable without video) and video-dependent QA (requiring video context). Evaluations show that SoTA MLLMs often hallucinate on video-dependent QA, relying on language priors instead of video inputs.

Foundational video generation models[14] have attracted significant attention due to their recent success in AI-Generated Content (AIGC). Their realistic and high-fidelity outputs open broad applications across multimedia[15][16], entertainment and content creativity[17], robotics and embodied AI[18]. Despite the

doi.org/10.32388/BXC6X1

promise of current video generation models, many critical flaws emerge during deployment. A key limitation lies in their failure to faithfully generate reality-grounded videos beyond following basic input prompt instructions and enhancing general video quality such as resolution[1]. Many state-of-the-art (SoTA) video generation models like Gemini-2.5-Pro[7], GPT-4o[19] and Qwen2.5-VL[8] (Figure. 1) exhibit hallucination issues such as inconsistent motion dynamics[20], and violations of real-world commonsense[21] and basic physics laws[22][23]. These shortcomings suggest that video generation models mainly optimize input-output alignment, reflected in the loss designs of CogVideo[24] and LaVie[25], without truly learning or applying real-world commonsense and physical laws. Instead, they tend to superficially imitate training data appearance and dynamics, rather than fundamentally understand the underlying principles of real-world behavior.

Detecting problematic generations make evaluating video generation more crucial, especially at scale[23][26]. Socre-based evaluation methods such as VideoScore[1] or DEVIL[27] produce a set of video quality scores without explanations, and they suffer from out-of-distribution data (newer released video generation models) and do not align with human visual judgments, as Figure. 1 suggests. One promising approach to facilitate a more interpretable video generation evaluation is by leveraging MLLMs for synthetic video understanding and evaluation. MLLMs offer insights into video understanding through their grasp of driving principles and commonsense knowledge, and can detect abnormalities by breaking down input prompts into subsets of questions and individually querying the video like FactScore[2] style, facilitating a more interpretable evaluation of video generation. However, current MLLMs are highly prone to hallucinations when evaluating synthetic videos. As shown in Figure. 1, even SoTA models hallucinate when presented with synthetic videos that are visually flawed and counter-intuitive, despite their strong performance on real-world videos[11][28]. We attribute this to the scarcity of high-quality and high-fidelity synthetic annotated videos in previous MLLMs' training data, as most are trained with real-world video data source[29][8][7][30]. Due to the lack of synthetic videos in the training data, MLLMs can only rely on their LLM backbones to reason, treating commonsense and physical knowledge as default ground truth assumptions without considering the actual visual input[31], which introduces biases stemming from the LLM's strong prior knowledge. While these biases may not surface on real-world videos, they cause severe hallucinations on synthetic videos, which often contain counterintuitive phenomenons that violate commonsense and physical knowledge (Figure. 1). Furthermore, although chain-of-thought reasoning often mitigates hallucinations in real-world understanding[32], the reasoning itself inherits deep-rooted hallucinations within MLLMs' language priors due to the scarcity of

annotated negative examples in synthetic video understanding tasks. Current MLLMs are unsuitable for reliably detecting commonsense violations in synthetic videos.

We propose VideoHallu, a benchmark consisting of expert-crafted, diverse question-answering pairs spanning context alignment, consistency, commonsense, and physical law reasoning of synthetically generated videos to bridge the gap between video generation and synthetic video evaluation. These reasoning-heavy tasks require an in-depth understanding of visual cues and language-based knowledge to answer our questions. We evaluate several SoTA video understanding models and provide a detailed analysis of their current failures on synthetic video understanding. Additionally, we use curriculum learning[33] and Group Relative Policy Optimization (GRPO)[13] to post-train an MLLM using datasets combining real-world (VideoLLaVA[34], PhysBench[22]) and synthetic videos (VideoHallu) to validate the effectiveness of improving reasoning abilities on synthetic video tasks through post-training. Our goal is to improve commonsense and physical law reasoning into synthetic video understanding. We hope VideoHallu and our RFT (Reinforcement Fine Tuning) framework offer insights into mitigating hallucinations and advancing video understanding and evaluation for synthetic videos. Our contributions include:

- We propose a novel benchmark VideoHallu, to test synthetic video understanding and evaluation tasks, featuring expert-annotated, diverse question-answer pairs covering alignment, spatial-temporal consistency, commonsense, and physical reasoning tasks.
- We conduct evaluation experiments on several state-of-the-art video understanding models using VideoHallu. Results show that the top-performing models achieve only a modest overall accuracy of 51.4%, with a highly uneven performance across perceptual-heavy and reasoning-heavy questions—exhibiting a gap of over 20% accuracy between the two.
- We use curriculum learning and GRPO to fine-tune QWen2.5-VL-7B using a small subset in combination of real-world and synthetic video understandings–Video-LLaVA-80K [34], Physbench [22], and VideoHallu. Our results show that even a limited amount of RFT synthetic data (800) yields a notable 3% overall accuracy improvement, with significant gains in commonsense and physics reasoning question accuracy, highlighting its strong potential for future exploration.

2. Related Work

2.1. Hallucination in Vision-language Models

Hallucinations in MLLMs—i.e., the tendency to generate outputs that are inconsistent with the target texts, images, or videos [35][36]—remain a long-standing challenge. Previous works have focused on developing benchmarks or pipelines to trigger hallucinations [31][37][38], as well as mitigating them through instruction tuning [39] or by enhancing the model's visual and textual representations [40]. Since hallucination arises from conflicts between the language priors of MLLMs and the actual visual inputs [38], it can be even 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[41] and EventHallusion[42], has aimed to establish benchmarks for evaluating model hallucination on both entities and events within videos, while also proposing methods to enhance the video understanding capabilities of MLLMs. HAVEN [43] further investigates the causes of hallucinations in video understanding models and introduces a video-thinking framework that incorporates reasoning and reflective thinking to mitigate hallucinations. 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.

2.2. Video Understanding Models

Video understanding is a fundamental task in computer vision [44][45]. Traditional video understanding [46][47] focuses on understanding the events happening over the timeline throughout the video. The emergence of foundation models in video understanding tasks [48][49] has enabled the development of general-purpose models capable of understanding videos across diverse domains, by leveraging large-scale training data and knowledge encoding through increased model capacity. Prior works mainly focus on using the instruction tuning framework [50] to unify the representation of video and language prompt into the same representation space [34][30] or scale up the visual-language models into the video domain [8][19][7]. Given the diverse deployment scenarios for video understanding models, recent works have focused on designing or distilling lightweight expert models for tasks such as video

question answering [51][52][53]. These models aim to enable applications on edge devices and have achieved competitive performance compared to large-scale models while using only a fraction of the parameters. Meanwhile, rule-based reinforcement learning has also been applied during the fine-tuning process [9][10] to encourage models to exhibit self-improvement and chain-of-thought reasoning abilities in video understanding. Despite the success of existing large models in real-world video understanding, their limitations—such as poor performance in probing commonsense and physical law violations in videos, as well as hallucinations in question answering—continue to hinder their broader adoption across diverse application scenarios. [54].

2.3. Evaluation of Video Generation Models

Video generation models[3][5][55][25][56][24] have a wide range of applications, including content creation, robotic environment simulation, and training[49]. However, the limited performance of video generation models—characterized by cross-modal misalignment[21] and frequent violations of commonsense[22]—remains a key bottleneck to their broader adoption and application. Many efforts have been devoted to designing comprehensive benchmarks for evaluating the quality of generated videos[57][58], as well as developing model judges[59] to assess the fidelity and coherence of generated content. Early explorations have primarily focused on incorporating human perception and preference annotations into the evaluation of video generation models[21][60], or on developing versatile evaluators that assess multiple dimensions of video quality, including spatial-temporal consistency, cross-modal alignment, and visual fidelity[1]. As video generation models continue to advance, recent efforts have shifted toward context-level evaluation of dynamic properties and actions, with a growing emphasis on assessing the representation of commonsense knowledge implicitly embedded in generation prompts[61], particularly those related to physical laws[62][63]. Notable investigations in this direction leverage human preference data to fine-tune evaluation models for detecting violations of physical laws arising from entity motion[23], or broader commonsense violations in layout generation and state transitions within large-scale scenes or dynamic entity interactions, often through the design of expert-informed metrics[64]. Despite substantial progress in this field, many open questions remain, as prior work has largely focused on quantitative evaluation of video quality rather than offering interpretable assessments or insights into entity-wise motion understanding and reasoning over evaluation outcomes—gaps that this work aims to address.

3. VideoHallu: Benchmark for Synthetic Video Understanding and Evaluation

3.1. Preliminary

Synthetic Videos

This work focuses on video understanding tasks using synthetic videos—i.e., videos generated by foundation models from textual prompts. Ideally, such videos should not only align with the prompts but also exhibit real-world plausibility, including: (1) smooth temporal changes in entities, (2) object geometry and dynamics consistent with common sense, and (3) entity motion that adheres to physical laws unless stated otherwise. We manually curate prompts to evaluate how well video generation models align with our predefined categories. For each prompt, we generate videos using seven SoTA models: Sora[4], Veo2[3], Kling[5], Runway Gen2[56], PixVerse[65], LaVie[25], and CogVideo[24].

An example prompt for our synthetic video generation is:

> Generate a scene of a feather and a heavy rock are released at the same height and begin to fall to the ground on Earth.

We keep prompts simple and specific to avoid confusing the generation models.

Synthetic Video Understanding and Evaluation

Despite growing interest in video generation, synthetic videos generated by current foundation models remain far from flawless. Visual abnormalities and counterintuitive phenomena are common, underscoring the need for models that can both evaluate video quality and detect such anomalies. MLLMs are emerging as prominent candidates for serving as comprehensive video judges. However, beyond basic quality assessment or scene description, synthetic video understanding requires models to: (1) detect, localize, and characterize anomalies; (2) apply commonsense and physical reasoning to assess plausibility; and (3) resist hallucinations when visual input conflicts with their language priors. This threefold challenge sets a higher standard for MLLMs, requiring them to approach human-level video perception to perform effectively.

3.2. Case study: Can MLLMs understand complex concepts as humans do in synthetic videos, and are current video evaluation models truly reliable?

A natural approach to using MLLMs as video judges, which can be done using either score-based models, VideoScore[1] and WorldScore[64] that assign scores to aspects like visual quality, alignment with prompts, and factual consistency, or prompt video understanding MLLMs with target-specific questions and performing answer matching or directly give a quality rating score[26]. To assess whether both kinds of models can handle synthetic video understanding tasks with awareness of anomalies, we present a case study as part of our assessment:

We first generate a synthetic video by prompting two video generation models, Veo2[3] and Runway Gen2[56], as follows

> A feather and a heavy rock are released into the air and start to fall.

As shown in Figure 1, the Veo2-generated video exhibits a commonsense violation, while the Runway Gen2 video is misaligned with the prompt and lacks physical consistencies. This case is designed to evaluate whether video judge models can resolve conflicts between their language priors and actual visual observations.

We first apply VideoScore[1] to assess both videos across metrics like visual quality, text-video alignment, and consistency. Both videos receive similar, modest scores across all metrics. However, the misalignment in the Runway Gen2 video is clearly noticeable, suggesting a gap in VideoScore's ability to detect such issues, which is consistent to conclusion[26] that MLLMs are still far from being as qualified video judges than humans do. Meanwhile, Veo2 video contains a commonsense violation not covered by VideoScore's evaluation scope, which explains its deceptively favorable score.

In the second stage, we conduct video-dependent question-answering using SoTA MLLMs on commonsense-related prompts, where models are required to answer based on video observations. Results show that all models—GPT-4o[19], Gemini-2.5-Pro[7], and Qwen-2.5-VL[8]—answered incorrectly (Figure 5), stating the rock reached the ground first, despite the clear visual evidence to the contrary. Moreover, when prompted with commonsense-only questions (i.e., without video context), all models gave the same response, indicating they defaulted to language priors. These results suggest that MLLMs are prone to hallucinations in synthetic video understanding and often rely on priors rather than actual visual input, unlike human perception.

This case study demonstrates that, unlike human perception, current MLLMs struggle to identify abnormalities in synthetic videos that conflict with real-world commonsense, even when the issues are perceptually obvious. The failures stem from limited video perception capabilities and conflicts between language priors and visual observations, often leading to hallucinations. To address this gap, a comprehensive benchmark focused on commonsense and physical reasoning in synthetic video understanding is essential for exploring and guiding MLLMs toward deeper, more reliable visual reasoning.

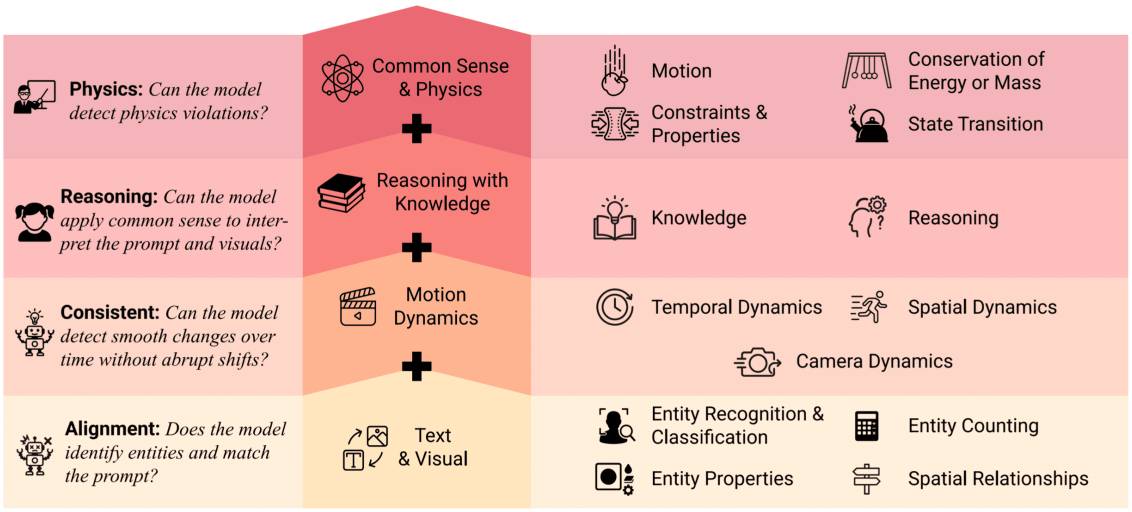


Figure 2. Question Categorization of VideoHallu. We design our benchmark, VideoHallu, with four question categories to probe hallucinations in synthetic video understanding, covering perceptual understanding to abstract reasoning. **(a) Alignment** checks if the model correctly identifies and understands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

3.3. Video Understanding and Evaluation Categorization

We draw inspiration from basic video quality evaluation definitions from MVBench[11] and WorldModelBench[23] 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 all question-answering pairs into four major categories with many sub-categories.

Alignment: This question 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: This question 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: This question 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 misunderstanding or 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: This question 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.

Our goal is to move beyond basic metrics like frame consistency or resolution and instead provide a deeper, more rigorous evaluation by identifying visual abnormalities across predefined categories. To this end, we craft targeted adversarial questions designed to explicitly reveal these anomalies. Our core motivation is to assess whether current SoTA MLLMs can effectively detect and interpret such abnormalities—an essential step toward scalable, interpretable video evaluation. We extend these video generation evaluation principles to form our video understanding criteria.

3.4. Data Collection

We first manually crafted 141 adversarial prompts that are challenging for video generation models to generate, including our four predefined categories. For each question-answering pair, each annotator assigns them a super category and a sub category. We gather five expert-level human annotators to craft 3233 question-answering pairs over generated videos, many of which are intentionally designed as reasoning-heavy questions to probe the counter-intuitive or misaligned with the input prompts to test whether video understanding can detect the abnormalities in these videos (Figure 4).[1] Our questions are meant to challenge current SoTA video generation models that requires understanding of the real world physics and commonsense reasoning abilities, and easily cause unusual phenomenons in the generated videos. For example, *the process of a glass breaking, or the process of a bullet shooting into a watermelon* are abstract concepts that are challenging for video generation and requiring complex understanding of the real world physics and commonsense reasoning abilities, and easily cause unusual phenomenons in the generated videos. However, being able to detect these abnormalities is a key component for facilitating more robust and interpretable video evaluation, but current SoTA MLLMs struggle at identifying them, which we study in detail in Section 4.

4. Experiment and Results

In this section, we evaluate 16 SoTA MLLMs on our curated test dataset (Table 1). For models not trained with reinforcement learning, we use standard prompting to elicit direct answers. For those trained with reinforcement learning or chain-of-thought supervised finetuning (e.g., Video-R1-cot[9] and VideoChat-R1-think[10]), we prompt them to reason step by step by analyzing key video frames before generating a final answer. Figure 4 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 are included in Appendix A.

Evaluation: LLM-as-a-judge[66][67][68] has shown promising improvement and high correlation with human judgments compared to previous work for simple and easy gold and generated answer comparison. Since our answers include free-form and open-eneded question-answer pairs, we use GPT-4-as-a-Judge to judge the correctness of model generated response with our written gold answers. We extract the final answer from Video-R1-cot and Videochat-R1-think as the generated response for GPT-4 to judge. To ensure the robustness and correctness of LLM-as-a-Judge in our dataset, we randomly sample 200 answer pairs with length greater than five words and annotate their correctness, with a 99.3% human-GPT agreements.

Model	Alignment				S-T Consistency			Commonsense		Physics				Overall
	A-EC	A-EP	A-ERAC	A-SR	SC-CD	SC-SD	SC-TD	CS-K	CS-R	P-C	P-CAP	P-M	P-ST	
MLLMs: < 7B														
SmolVLM-3B[53]	17.3	9.9	12.1	7.1	11.9	12.0	17.3	6.6	13.0	9.5	0.0	12.2	0.0	13.4
InternVL3-2B[69]	44.3	49.3	58.0	26.2	16.7	32.4	35.0	44.3	31.9	33.3	23.5	37.8	30.0	39.9
Qwen2.5-VL-3B[8]	44.3	49.8	59.8	40.5	40.5	33.1	37.7	42.6	42.0	38.1	23.5	40.0	60.0	**42.9**
MLLMs														
Video-LLaVA[34]	44.9	40.4	48.3	21.4	38.1	29.6	33.7	37.7	30.4	**57.1**	17.6	41.1	20.0	37.5
LLaVA-NeXT[70]	44.9	53.7	50.6	26.2	31.0	35.2	33.0	36.1	36.2	**57.1**	29.4	26.7	0.0	39.1
Video-LLaMA[30]	53.5	51.2	56.3	42.9	54.8	43.0	38.6	42.6	34.8	42.9	17.6	38.9	50.0	45.0
InternVL3-9B	45.9	55.7	58.0	40.5	40.5	37.3	42.9	50.8	39.1	47.6	29.4	46.7	50.0	46.4
InternVL3-38B	52.4	57.1	54.6	45.2	40.5	38.0	42.9	50.8	40.6	23.8	41.2	38.9	60.0	46.6
InternVL3-14B	49.2	53.2	57.5	50.0	42.9	37.3	42.9	44.3	40.6	38.1	47.1	47.8	60.0	46.7
Qwen2.5-VL-7B	53.5	**63.1**	65.5	54.8	47.6	42.3	**43.9**	50.8	42.0	42.9	**58.8**	45.6	40.0	51.0
Qwen2.5-VL-32B	54.6	60.1	**67.8**	52.4	**59.5**	42.3	40.7	55.7	**58.0**	42.9	52.9	**51.1**	40.0	**51.4**
MLLMs: R1-finetuned														
VideoChat-R1[10]	47.0	55.2	63.2	38.1	42.9	34.5	32.2	65.6	39.1	42.9	52.9	44.4	60.0	44.2

doi.org/10.32388/BXC6X1

Model	Alignment				S-T Consistency			Commonsense		Physics				Overall
	A-EC	A-EP	A-ERAC	A-SR	SC-CD	SC-SD	SC-TD	CS-K	CS-R	P-C	P-CAP	P-M	P-ST	
VideoChat-R1-think	47.0	56.7	63.2	42.9	42.9	34.5	34.1	**67.2**	40.6	52.4	47.1	43.3	40.0	45.3
Video-R1-CoT[9]	55.1	55.2	63.2	42.9	40.5	**47.2**	39.2	55.7	46.4	28.6	35.3	47.8	40.0	48.3
Video-R1-SFT-CoT	53.5	55.7	68.4	38.1	45.2	45.1	40.3	45.9	47.8	33.3	41.2	50.0	50.0	50.6
Video-R1-SFT	50.3	60.6	69.0	40.5	47.6	43.7	42.6	57.4	46.4	38.1	52.9	48.9	**60.0**	50.6
Video-R1	51.4	62.1	**67.8**	35.7	40.5	45.1	43.7	52.5	46.4	38.1	**58.8**	50.0	**60.0**	**50.8**
MLLMs: Black-Box														
GPT-4o[6]	45.4	58.1	61.5	45.2	40.5	35.9	37.3	54.1	34.8	47.6	47.1	46.7	50.0	45.5
Gemini-2.5-Flash[7]	54.1	60.0	65.0	60.0	49.2	55.1	42.9	47.1	37.8	40.0	45.2	47.2	41.4	49.6
Gemini-2.5-Pro[7]	**56.8**	61.6	65.5	**57.1**	50.0	41.5	40.9	52.5	46.4	38.1	47.1	35.6	40.0	**49.8**

Table 1. SoTA MLLM Evaluation on VideoHallu. We evaluate diverse SoTA models across sizes and training strategies, reporting both overall and sub-category accuracies. Qwen2.5-VL-32B achieves the highest overall performance among all models.

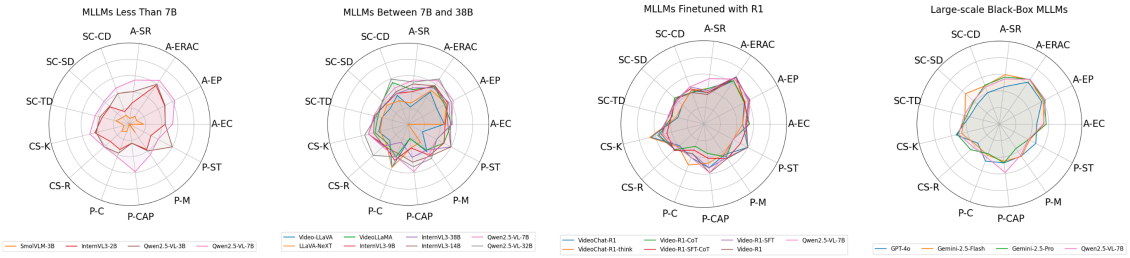


Figure 3. SoTA MLLM Evaluation on VideoHallu Across Sub-Categories. We evaluate SoTA MLLMs on VideoHallu, with results broken down by sub-category. From left to right, we show: (a) models under 7B parameters; (b) models between 7B–38B; (c) R1 fine-tuned models; and (d) large black-box MLLMs. While many perform well on alignment tasks, they remain prone to hallucinations in reasoning-heavy tasks, with notably weaker performance on physics and commonsense reasoning.

4.1. Results and Analysis

MLLMs struggle to detect counterintuitive phenomena and abnormalities in generated videos.

Qwen-VL series (7B/32B) achieves the highest overall accuracy at around 51% (Table 1). However, even the best models often fail to answer simple visual questions about synthetic videos, relying heavily on language priors or prior knowledge from real-world videos, which leads to unfaithful or incorrect assumptions from real-life data (Figure 5). Qwen2.5-VL's advantage stems from broader training dataset coverage, a vision encoder tailored for video tasks, and strong cross-modal alignment. In contrast, models like GPT-4o treat video inputs as frame sets without modeling temporal dynamics, falling behind other competitive SoTA models—Gemini-2.5-pro and Gemini-2.5-Flash.

Despite Qwen2.5-VL's lead, the overall accuracy remains below 60% across all models. The performance across sub-categories is uneven: models excel at tasks requiring direct observation (e.g., instruction-following, alignment, spatial-temporal consistency), but struggle with reasoning-intensive questions involving common sense or physics knowledge. Within each category, perception-oriented tasks like entity recognition yield higher accuracy than reasoning-heavy ones like spatial relations or entity counting, where language priors and visual priors from real-life videos often introduce hallucinations. This highlights current models' limited capacity for integrating memorized knowledge into multi-modal reasoning.

doi.org/10.32388/BXC6X1

Larger models tend to achieve higher accuracy, but their performance is ultimately constrained by the capabilities of the vision encoder.

We evaluate models across a range of sizes, from small open-source models like InternVL3-2B to large closed-source ones like Gemini-2.5-Pro and GPT-4o. Generally, larger models demonstrate stronger language and visual reasoning abilities on all categories Our results confirm this trend: within the same series, such as Qwen and InternVL series, larger models outperform smaller ones across tasks. However, for models within the same series (Qwen2.5-VL, InterVL3), accuracy gains reach peak on synthetic videos once the model size exceeds 7B, showing that further improvements in video understanding rely less on scaling the language model and more on enhancing the vision encoder and the quality of training data. In addition, overall performance is also highly relevant to how models perceive video inputs. GPT-4o tends to treat videos frame-by-frame, aiding frame-level alignment checks but harming inter-frame motion tracking, which again shows the importance of vision encoder architecture in video understanding.

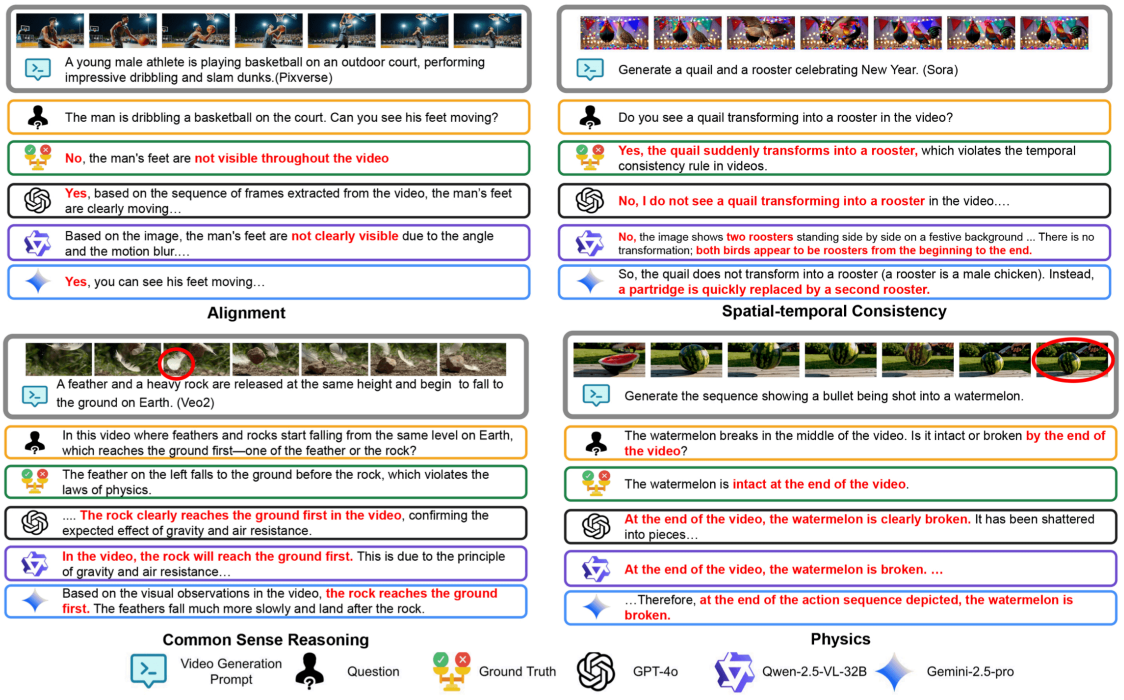


Figure 4. Hallucination showcases for SoTA models on VideoHallu. We collect hallucination cases observed during SoTA MLLM evaluations on synthetic video tasks. Each example includes the generation prompt, key frames, questions, human-annotated ground truth, and hallucinated answers from GPT-4o, Qwen2.5-VL, and Gemini-2.5-Pro, with hallucinations marked in Red. Video examples are available (here).

doi.org/10.32388/BXC6X1

4.2. Does Chain-of-Thought Reinforcement Learning Help Synthetic Video Understanding?

Chain-of-thought reasoning encourages models to rely more on prior language knowledge, often at the expense of visual grounding, which increases the risk of hallucinations.

Despite the recent success of the DeepSeek series[13], where Reinforcement Fine-Tuning (RFT) methods like GRPO[71] have enhanced reasoning in small models like Qwen2-VL-7B on math problems, our results show limitations of chain-of-though RL algorithms. In Table 1, we evaluate two RFT-trained video understanding models, Video-R1[9] and VideoChat-R1[10], under different prompt styles: short-answer to directly generate the answers (Video-R1, VideoChat-R1) and chain-of-thought to first think step by step then generate the final answer (Video-R1-CoT, VideoChat-R1-think). Both RFT models underperform their base model (Qwen2.5-VL-7B), and short-answer prompt style outperform those explicitly prompted for reasoning. This suggests that current RFT approaches, often based on limited, homogeneous data, boost performance on in-distribution tasks (e.g., real-world video understanding) but fail to generalize to synthetic videos. While chain-of-thought RFT can enhance performance on reasoning-heavy tasks like math problems, it is less effective for synthetic video understanding, where videos often depict counterintuitive or abnormal phenomena that conflict with real-life phenomenons. In such cases, encouraging extensive reasoning can cause the LLM backbone to rely too heavily on real-world commonsense knowledge, overlooking synthetic visual evidence and leading to hallucinated responses. GRPO-based RFT further amplifies this issue by encouraging models to develop their own reasoning patterns using weak supervision and non-semantic rewards like ROUGE[72], which poorly correlate with human judgment[73][68]. This misalignment creates a "perception gap" between model and human reasoning, often resulting in hallucinations—particularly on reasoning-heavy questions—explaining the weaker performance of chain-of-thought RFT models on our benchmark.

4.3. Can MLLMs learn counter-intuitive commonsense knowledge from synthetic data and curriculum learning?

Although Video-R1 is trained on massive number of diverse datasets, it still falters on synthetic video understanding, or even show lower to no improvement on VideoHallu. In this section, we use curriculum learning[33] to train Qwen2.5-VL-7B on three datasets in the order of difficulty– 4,500 from the video QA subsection of Video-LLaVA, 1,000 from PhysBench[22], and 800 from VideoHallu. We aim to study two

research questions: 1. *Can we improve MLLMs' reasoning and understanding abilities on synthetic videos through reinforcement learning finetuning?* 2. *Does curriculum learning help video understanding?* We combine the training dataset for GRPO training from the following three sources: One partition of VideoHallu dataset, the subsection of the open-end question-answering dataset of Video-LLaMA, along with the videos with physical law knowledge from PhysBench[22]. The combination of the fine-tuning dataset is inspired by the fine-tuning procedure of Video-R1[9], where we try to improve the models' video understanding abilities over common sense and physical laws by designing the entire training process in a curriculum manner with three datasets.

General-Real-World (GRW): The dataset for real-world video question-answering (Video-LLaMA[34]), to improve the base models' ability in video understanding. We sample 3,000 videos from this dataset as the training set.

Physics-Real-World (PRW): PhysBench[22] focuses on real-world common and physical knowledge, while this stage is to assist the model to understanding the indications and procedures, further refine the representations of physical and common sense-related knowledge, with the language prior within the video understanding model. We sample 1,000 videos from the dataset as the training set.

Synthetic Reasoning (SR): The dataset contains all synthetic videos question-answering (VideoHallu), focusing on common and physical knowledge. This stage helps the model overcome potential hallucinations within the language priors when reasoning over synthetic video and become more sensitive in probing common sense and physics violations in synthetic videos. Our training set size is 800 video QA pairs.

Specifically, we use GRPO to train Qwen2.5-VL-7B on three datasets separately first– GRW-R1-7B, PRW-R1-7B, SR-R1-7B. We use Answer-Equivalence BERT[68], which has been fine-tuned specifically on answer-correctness pairs, to evaluate the correctness of generated answers compared to the gold answer. The reward is defined as the similarity score between the embedding representations of the predicted answer a_{pred} and the gold answer a_{gold}, calculated similarly to BERTScore[74] as follows:

$$\text{Reward}(a_{\text{pred}}, a_{\text{gold}}) = \frac{\mathbf{E}(a_{\text{pred}})^\top \mathbf{E}(a_{\text{gold}})}{\|\mathbf{E}(a_{\text{pred}})\| \cdot \|\mathbf{E}(a_{\text{gold}})\|}, \tag{1}$$

where $E(a_{gold})$ and $E(a_{pred})$ denote the embedding vectors of the predicted and gold answers from the finetuned BERT, respectively. We train each model group with one epoch with gradient accumulation step of 1. We use the saved checkpoints for GRW and PRW and then train the model on the synthetic dataset to

perform curriculum learning. Our results (Table 2) show that only training on general-world data does not improve model's understanding abilities on synthetic video understanding, but adding physics knowledge can improve synthetic video understanding robustness mainly in the physics domain. In addition, we show that curriculum learning to first learn general physics knowledge from the data, then from synthetic videos can help models gain most understanding on synthetic video understanding.

Model	Alignment				S-T Consistency			Commonsense		Physics				Overall
	A-EC	A-EP	A-ERAC	A-SR	SC-CD	SC-SD	SC-TD	CS-K	CS-R	P-C	P-CAP	P-M	P-ST	
MLLMs: Previous SoTA														
GPT-4o	45.4	58.1	61.5	45.2	40.5	35.9	37.3	54.1	34.8	47.6	47.1	46.7	50.0	45.5
InternVL3-14B	49.2	53.2	57.5	50.0	42.9	37.3	42.9	44.3	40.6	38.1	47.1	47.8	60.0	46.7
Gemini-2.5-Pro	**56.8**	61.6	65.5	**57.1**	50.0	41.5	40.9	52.5	46.4	38.1	47.1	35.6	40.0	49.8
Video-R1	51.4	62.1	67.8	35.7	40.5	45.1	43.7	52.5	46.4	38.1	58.8	50.0	60.0	50.8
Qwen2.5-VL-7B	53.5	**63.1**	65.5	54.8	47.6	42.3	43.9	50.8	42.0	42.9	58.8	45.6	40.0	51.0
Qwen2.5-VL-32B	54.6	60.1	67.8	52.4	**59.5**	42.3	40.7	55.7	58.0	42.9	52.9	**51.1**	40.0	**51.4**
Training Separately														
GRW-R1-7B	48.7	60.3	67.5	34.9	51.7	49.1	42.9	57.7	60.0	62.2	56.9	42.7	50.0	51.5
PRW-R1-7B	49.3	60.3	67.1	41.9	55.2	43.9	57.1	57.7	60.0	**64.9**	54.9	41.3	70.0	52.2
SR-R1-7B	50.7	58.7	68.3	48.8	56.9	**50.9**	**66.7**	61.5	61.4	54.1	56.9	43.4	70.0	**53.4**
Curriculum Learning														
GRW+SR-R1-7B	51.3	59.8	**68.9**	51.2	53.5	42.1	47.6	57.7	56.4	56.8	**60.0**	46.2	60.0	52.1
PRW+SR-R1-7B	50.7	58.7	67.7	53.5	55.2	47.4	62.0	**61.6**	**62.1**	56.8	56.9	44.8	80.0	**54.2**

Table 2. Fine-Tuned Model Evaluation on VideoHallu. We evaluate models fine-tuned on either domain-specific sub-datasets or curriculum-based composite datasets. Results show that models trained only on general real-world videos yield little to no gains on synthetic video understanding. Incorporating general

doi.org/10.32388/BXC6X1

physics data improves physics reasoning, and a curriculum starting with real-world physics followed by synthetic data leads to a 2.8% performance boost.

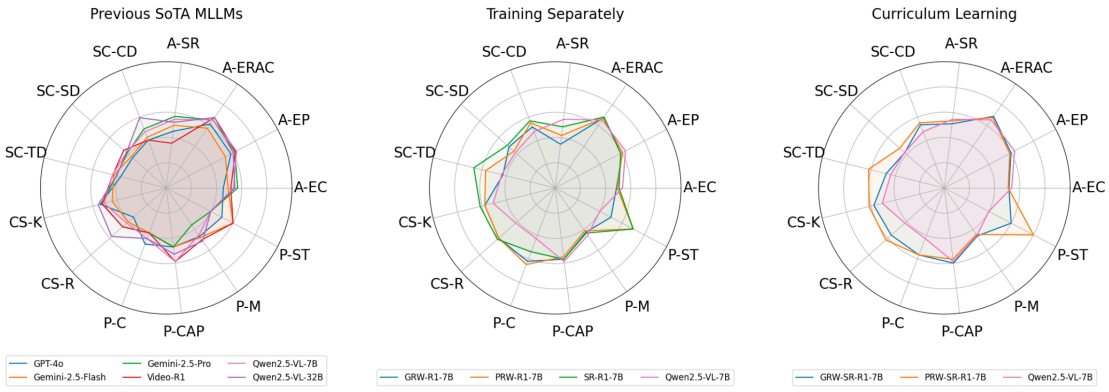


Figure 5. Evaluation Breakdown of Fine-Tuned Models. We show results for (a) previous SoTA MLLMs, (b) models fine-tuned on sub-datasets, and (c) models fine-tuned on the full dataset via curriculum learning. Compared to the baseline (Qwen2.5-VL-7B), reinforcement fine-tuning on commonsense and physics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

4.4. Discussions

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

1. Visual encoders constrain MLLMs' performance in synthetic video understanding.

In Section 4.1, all MLLMs evaluated suffer from hallucination in synthetic video tasks, achieving poor overall accuracy below 52% and consistently performing better on perception-oriented tasks (e.g., alignment, spatial-temporal consistency) than on reasoning-heavy tasks involving commonsense and physical reasoning. This gap partly stems from how MLLMs perceive videos: current visual encoders struggle to capture context-level temporal and spatial dynamics. Instead of flexibly encoding entity-level, semantic-sensitive information critical for synthetic video understanding, MLLMs tokenize videos into patches or fixed-length clips, inherited from MLLMs' visual encoders, which limits their ability to handle the non-realistic, dynamic patterns common in synthetic videos.

doi.org/10.32388/BXC6X1

2. Hallucination in synthetic video understanding stems from language priors, which makes detecting synthetic visual abnormalities worse.

We observe severe hallucinations in the reasoning processes of synthetic video understanding. SoTA models struggle at reasoning-oriented QA tasks, and RFT only on real-world data even worsens performance (Table 1). This stems from current MLLMs' strong language priors, learned from datasets dominated by real-world videos with explicitly represented contexts, where commonsense and physical laws are typically implicit. Despite seeing annotated synthetic examples during training, models assume these principles as default truths without truly learning from visual cues. As a result, when faced with generated videos that contain content conflicting with real-world phenomena, models struggle to reconcile the visual input with their prior knowledge. Furthermore, reasoning built on these flawed understandings leads to even more severe hallucinations, as shown in Video-R1-cot and VideoChat-R1-think (Table 1). We propose that future MLLMs for synthetic or physical-reasoning video tasks should focus on refining reasoning abilities to better leverage language priors while actively correcting hallucinated vidual understandings.

3. Both high-quality negative examples and RFT matter.

In Section 4.2, we show that RFT-enhanced models like Video-R1 and VideoChat-R1 do not outperform their base models (Qwen2.5-VL-7B) on synthetic video understanding tasks. However, in Section 4.3, we demonstrate that reinforcement fine-tuning methods like GRPO, when combined with a curated dataset covering both general and physics-specific video understanding across real and synthetic videos can improve model performance. Our results reveal that it is the quality and coverage of the data—not just the fine-tuning method—that drive improvements. With both positive and negative examples, clearly annotated reasoning procedures, and reasoning-stimulating post-training like GRPO, even small-scale models like Qwen2.5-VL-7B achieve QA accuracy gains on validation sets. We show the importance of combining high-quality data and reasoning-focused fine-tuning to enhance synthetic video understanding in future models, encouraging models to fundamentally grasp commonsense and physics knowledge and apply them in reasoning about more advanced concepts.

5. Conclusion

We introduce VideoHallu, a novel benchmark targeting hallucination probing and mitigation in synthetic video understanding. We curate expert-annotated, diverse, reasoning-heavy QA pairs covering

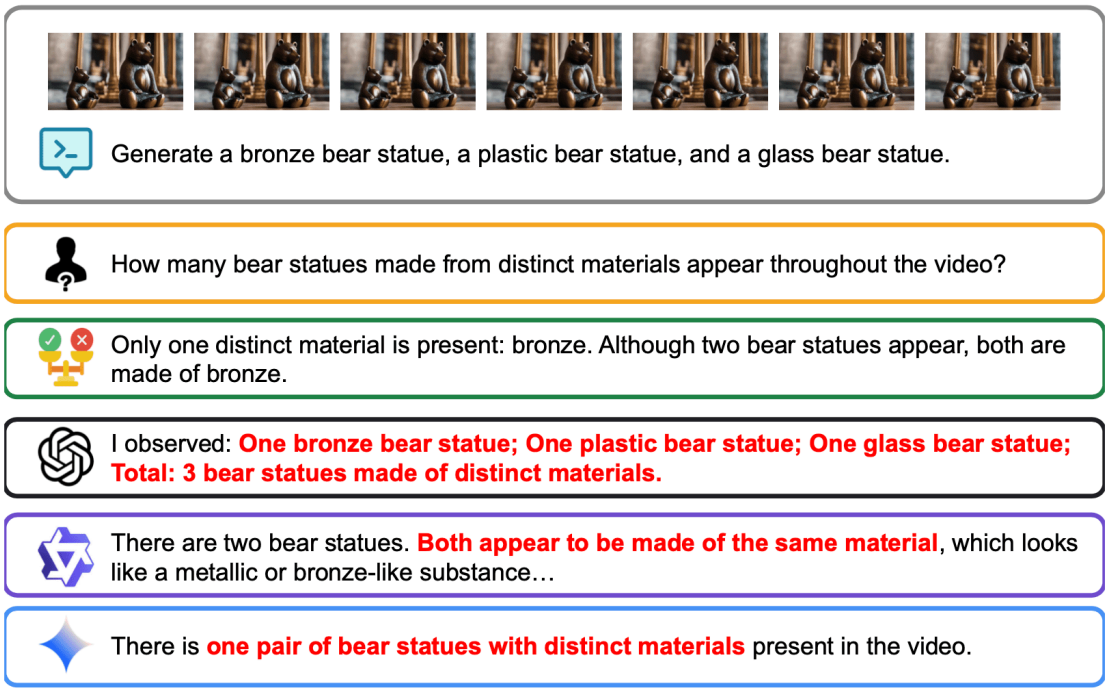
alignment, spatial-temporal consistency, commonsense, and physics reasoning to probe hallucinations induced by real-world training priors in MLLMs that are not generalized to synthetic generated videos. Evaluation on SoTA MLLMs show hallucination and poor performance on synthetic videos. We fine-tune MLLMs using GRPO with real-world video data, physics reasoning tasks, and VideoHallu synthetic videos organized via curriculum learning, achieving accuracy improvements. Our results show the importance of incorporating physics and commonsense reasoning data when refining MLLMs for synthetic video tasks. However, scalability remains a limitation, as producing high-quality annotations is costly. Future work will focus on expanding the dataset from synthetic videos and enhancing reinforcement fine-tuning to foster stronger and more robust reasoning in MLLMs, ultimately advancing solid physical-world understanding in video tasks.

Role	Content
System	You are a helpful assistant for video understanding. Answer questions based only on the visual content.
User	Given the video and the following question, select the most accurate answer choice. Explain your reasoning step by step before selecting the final answer.
Video Description	A feather and a rock are dropped from the same height on Earth. Both fall to the ground.
Question	Which object hits the ground first?
Options	A. The feather B. The rock C. Both hit the ground at the same time
Answer	[LLM will complete this section with reasoning and answer]

Table 3. Example prompt structure used for querying the LLM in our benchmark.

Appendix A Hallucination Showcases

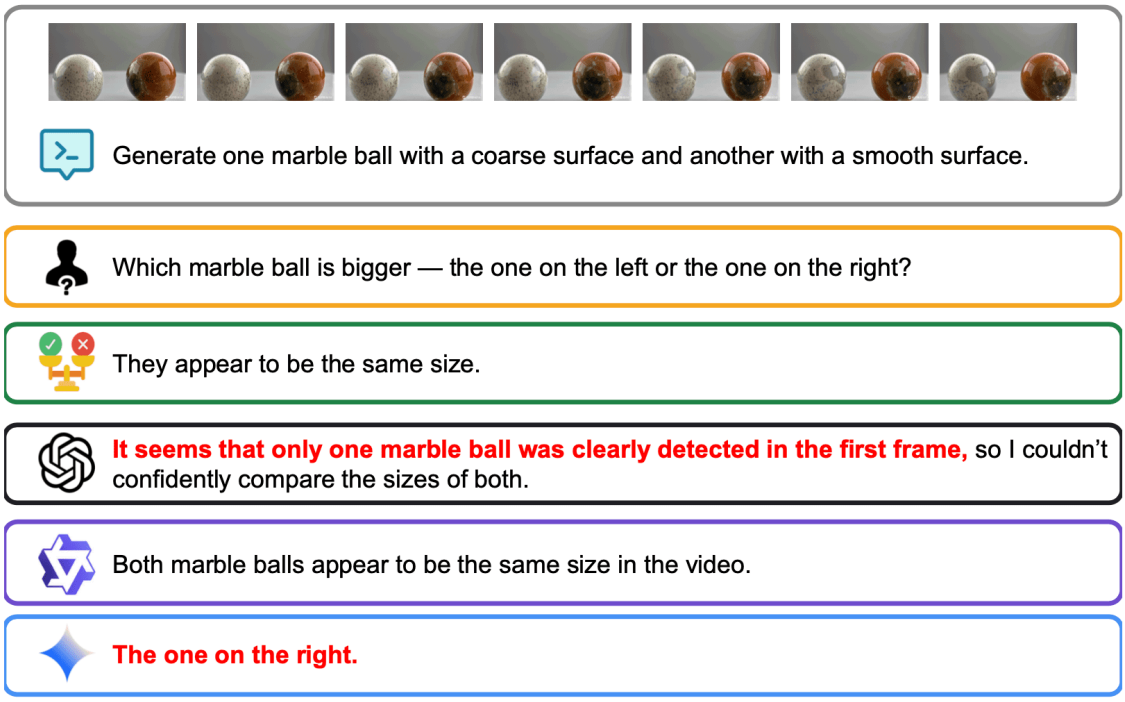
We present selected cases from SoTA MLLM evaluations across each VideoHallu sub-categories. Hallucinations in model answers, common sense or physics violations in videos, and other notable cues in the video, questions, or ground truth are highlighted to assist the reader's understanding.



Alignment - Entity Counting

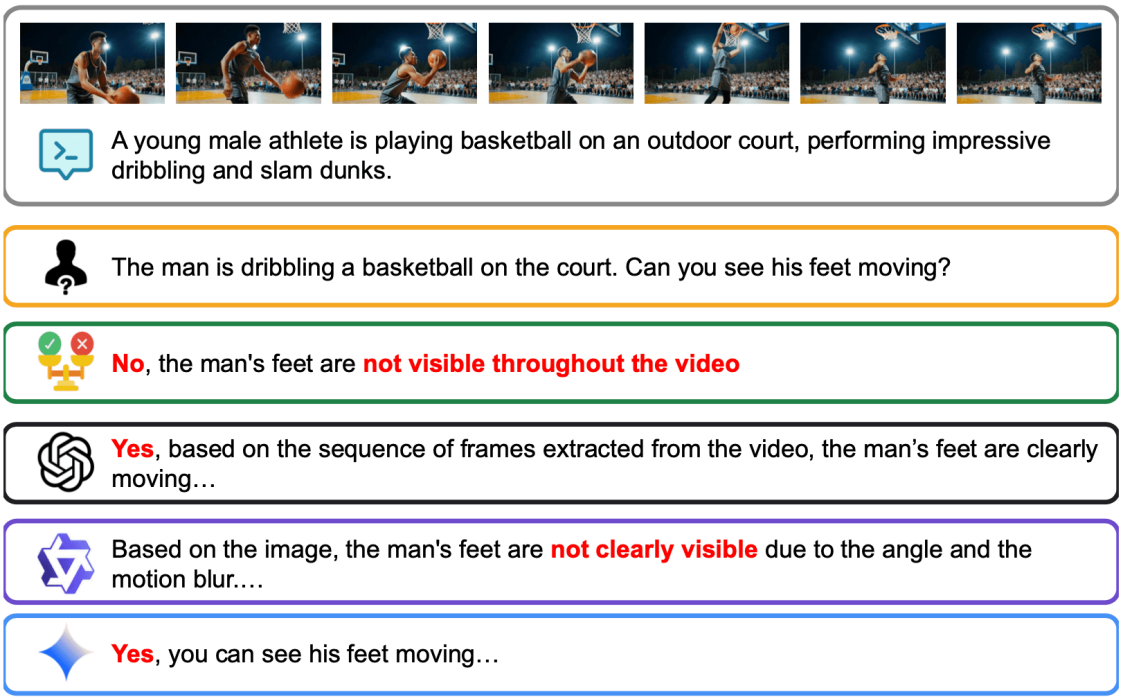
Figure 6. Hallucination Case from Alignment – Entity Counting (A-EC). We show hallucination examples from SoTA MLLM evaluations under the A-EC category. Each case includes the video generation prompt (Gray), key frames from synthetic videos (Gray), questions (Orange), ground truth (Green), and model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

doi.org/10.32388/BXC6X1



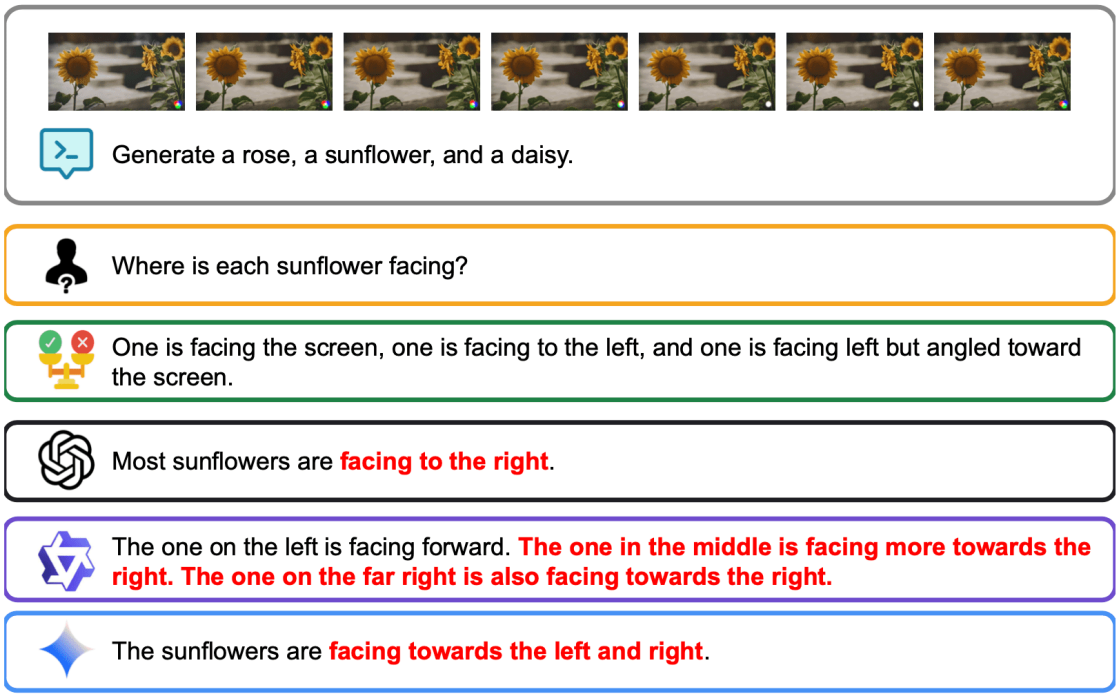
Alignment - Entity Properties

Figure 7. 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.



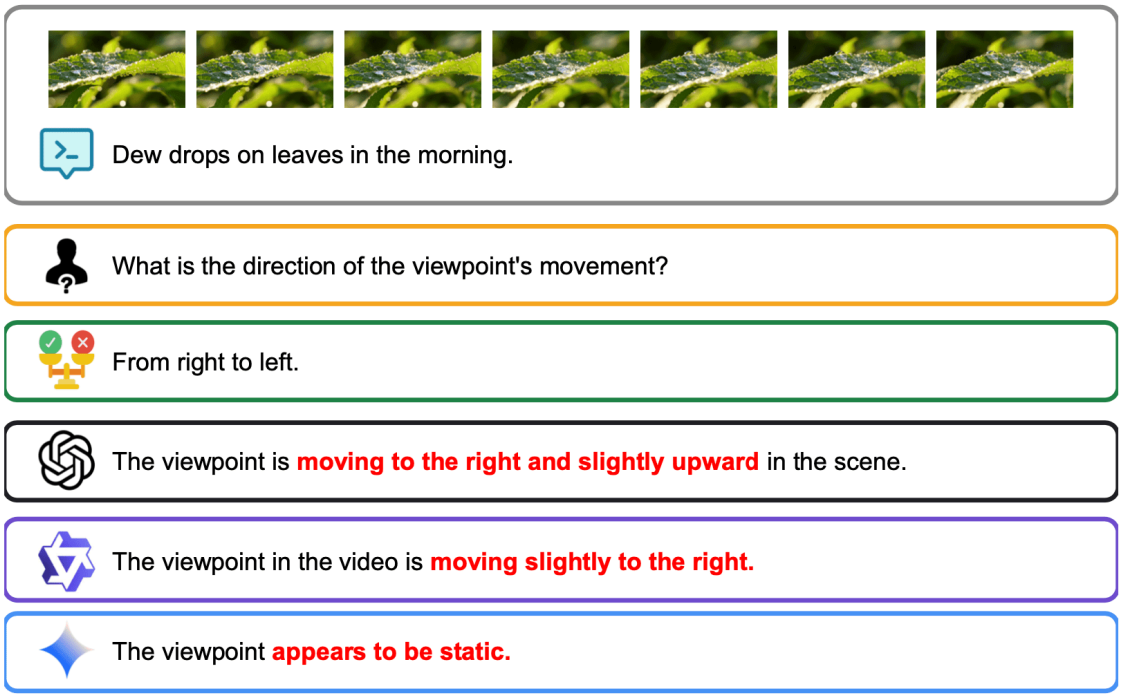
Alignment - Entity Recognition and Classification

Figure 8. 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.



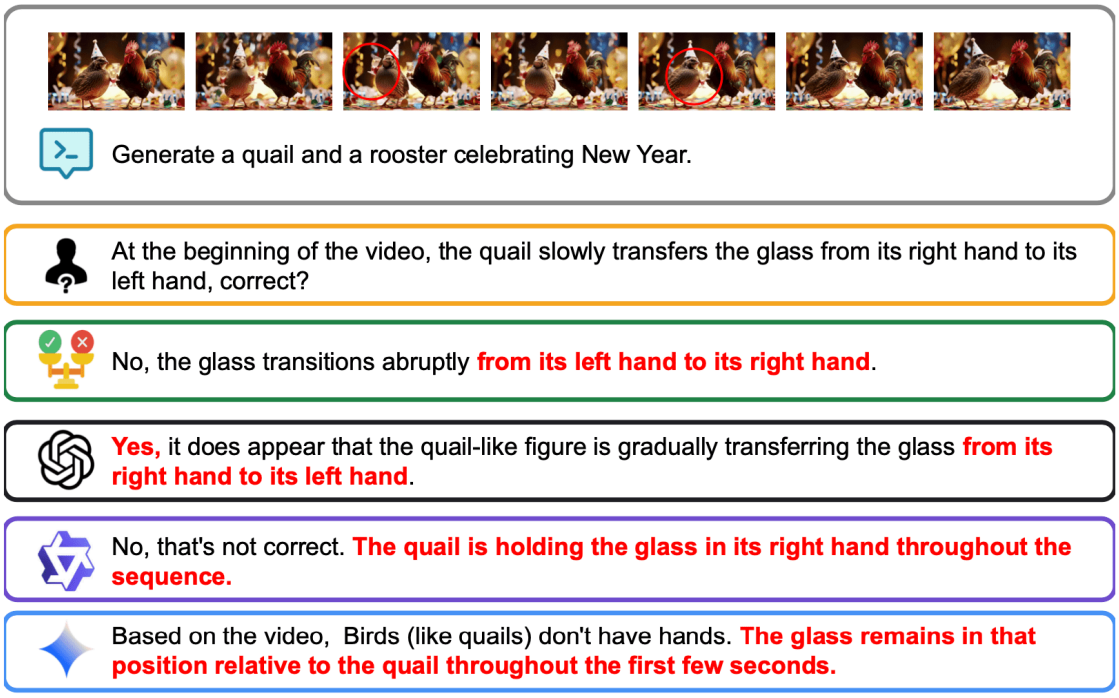
Alignment - Spatial Relationships

Figure 9.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.



Spatial-temporal Consistency - Camera Dynamics

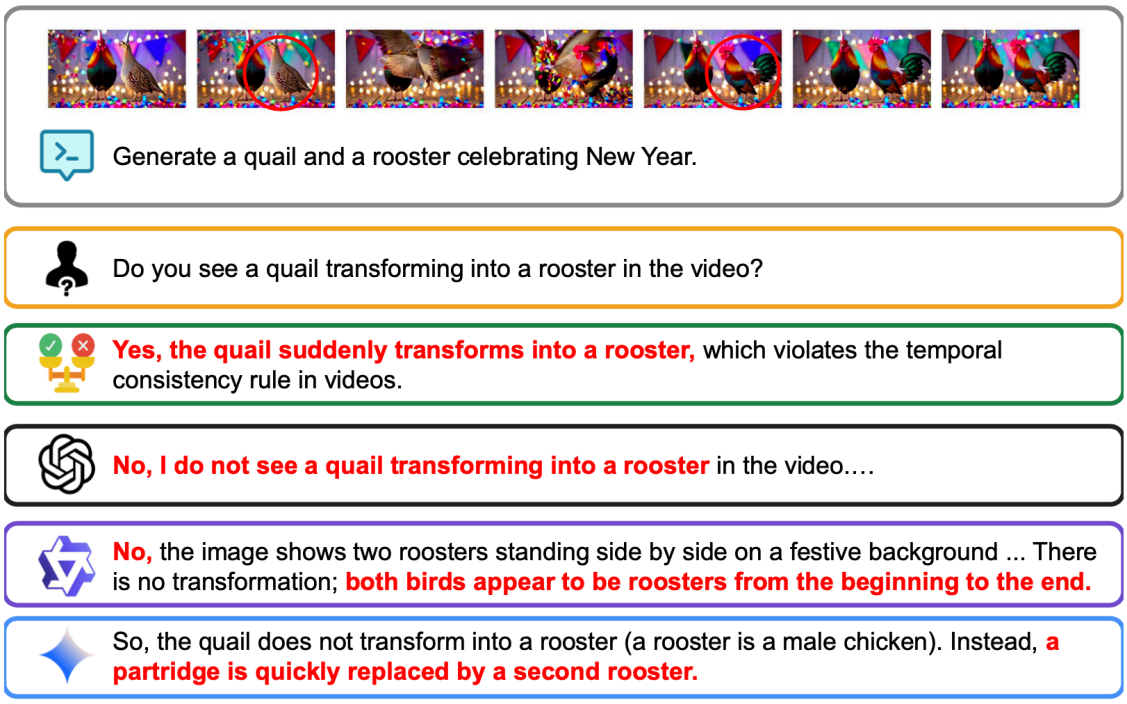
Figure 10. Hallucination Case from Spatial-temporal Consistency – Camera Dynamics (SC-CD). We show hallucination examples from SoTA MLLM evaluations under the SC-TD category. Each case includes the video generation prompt (Gray), key frames from synthetic videos (Gray), questions (Orange), ground truth (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.



Spatial-temporal Consistency - Spatial Dynamics

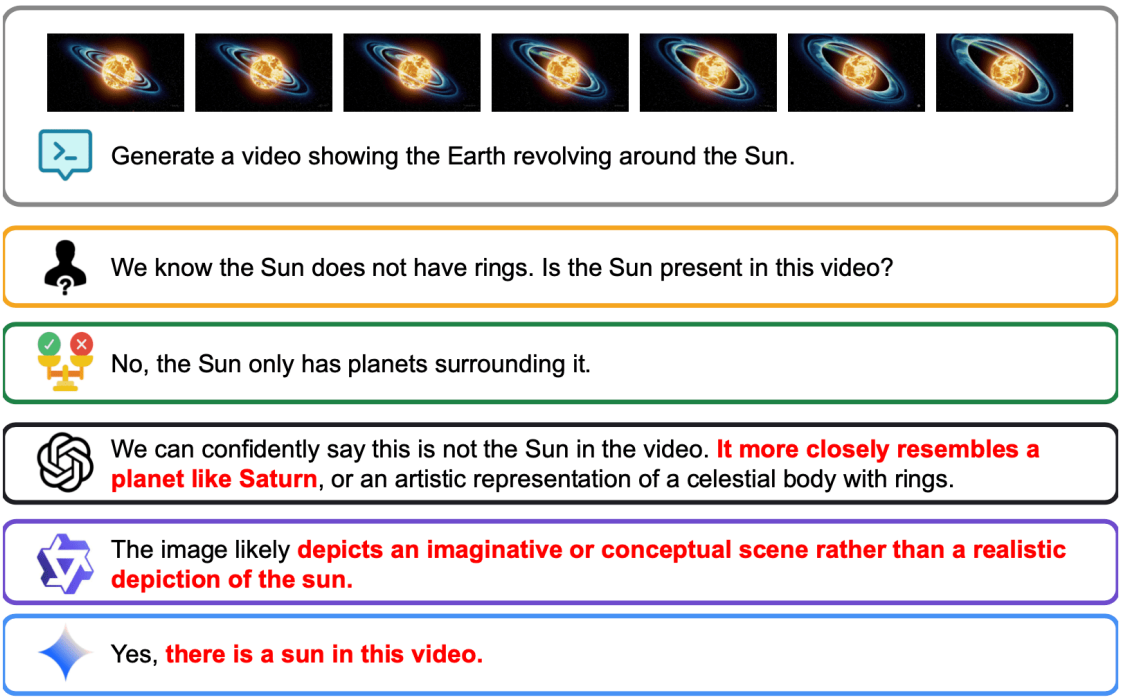
Figure 11. Hallucination Case from Spatial-temporal Consistency – Spatial Dynamics (SC-SD). We show hallucination examples from SoTA MLLM evaluations under the SC-SD category. Each case includes the video generation prompt (Gray), key frames from synthetic videos (Gray), questions (Orange), ground truth (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.

doi.org/10.32388/BXC6X1



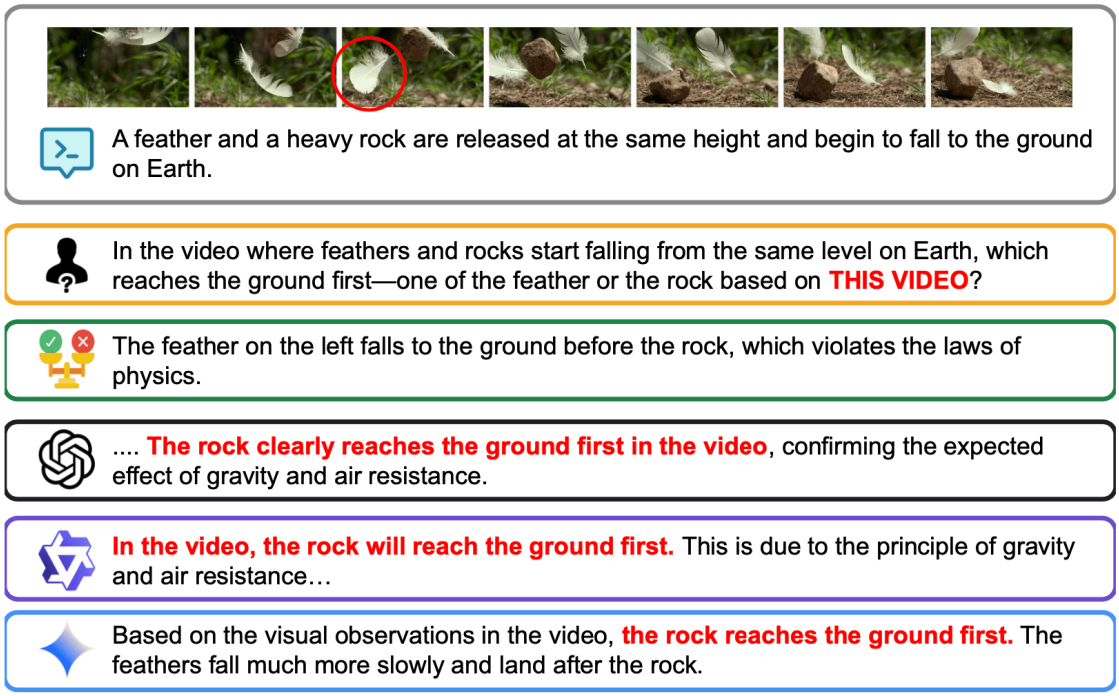
Spatial-temporal Consistency - Temporal Dynamics

Figure 12. 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.



Common Sense Reasoning - Knowledge

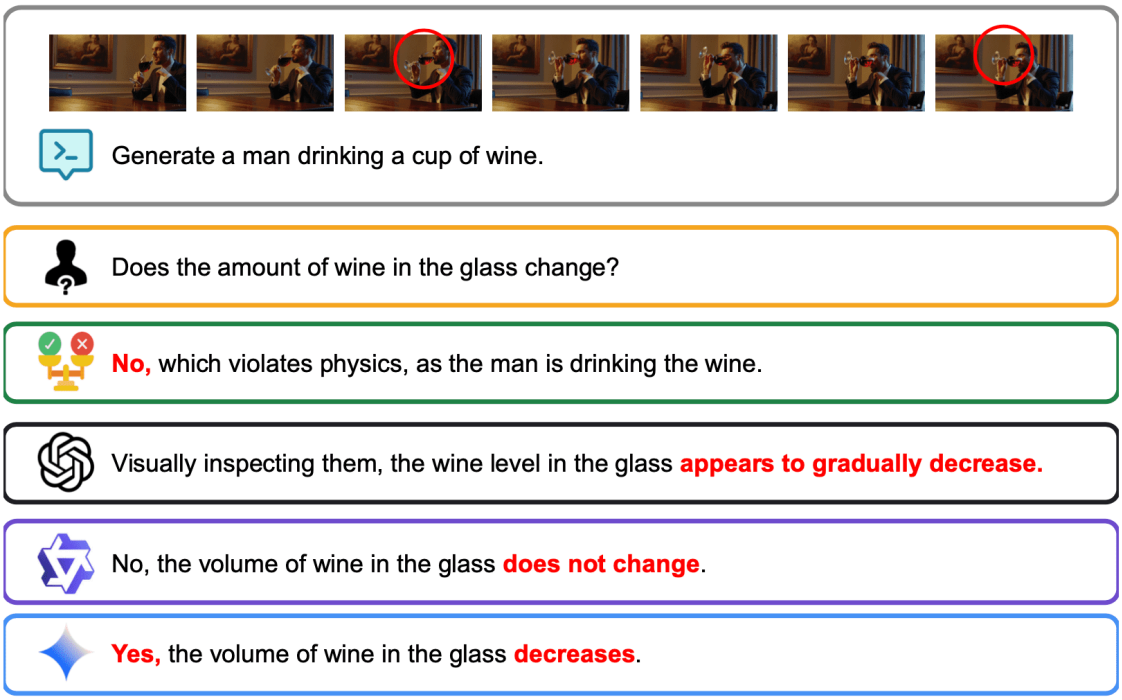
Figure 13. Hallucination Case from Common Sense Reasoning - Knowledge (CS-K). We show hallucination examples from SoTA MLLM evaluations under the CS-K category. Each case includes the video generation prompt (Gray), key frames from synthetic videos (Gray), questions (Orange), ground truth (Green), and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



Common Sense Reasoning - Reasoning

Figure 14. 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.

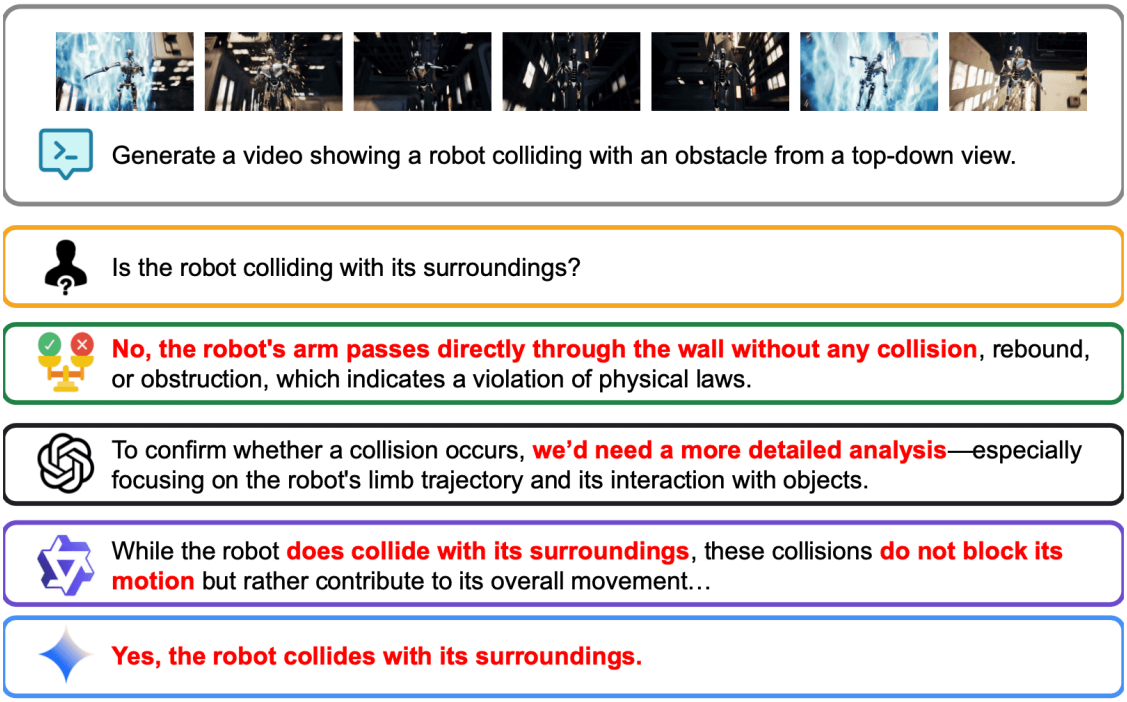
doi.org/10.32388/BXC6X1



Physics - Conservation

Figure 15. 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.

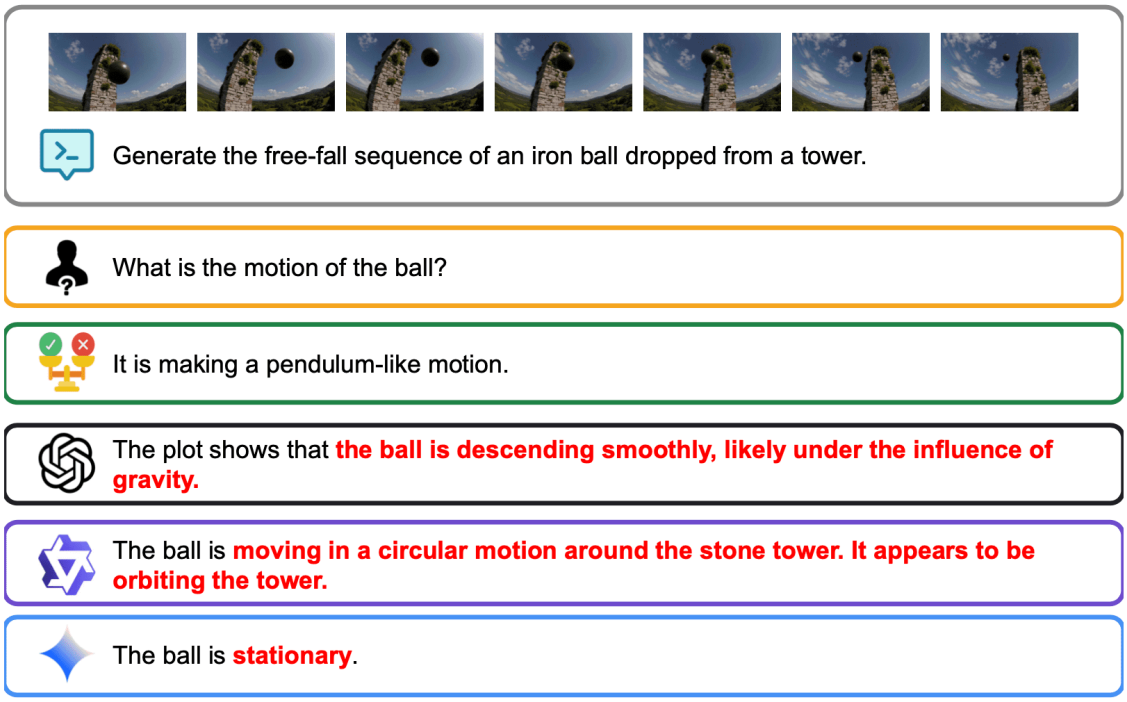
doi.org/10.32388/BXC6X1



Physics - Constraints and Properties

Figure 16. 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.

doi.org/10.32388/BXC6X1



Physics - Motion

Figure 17. 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.

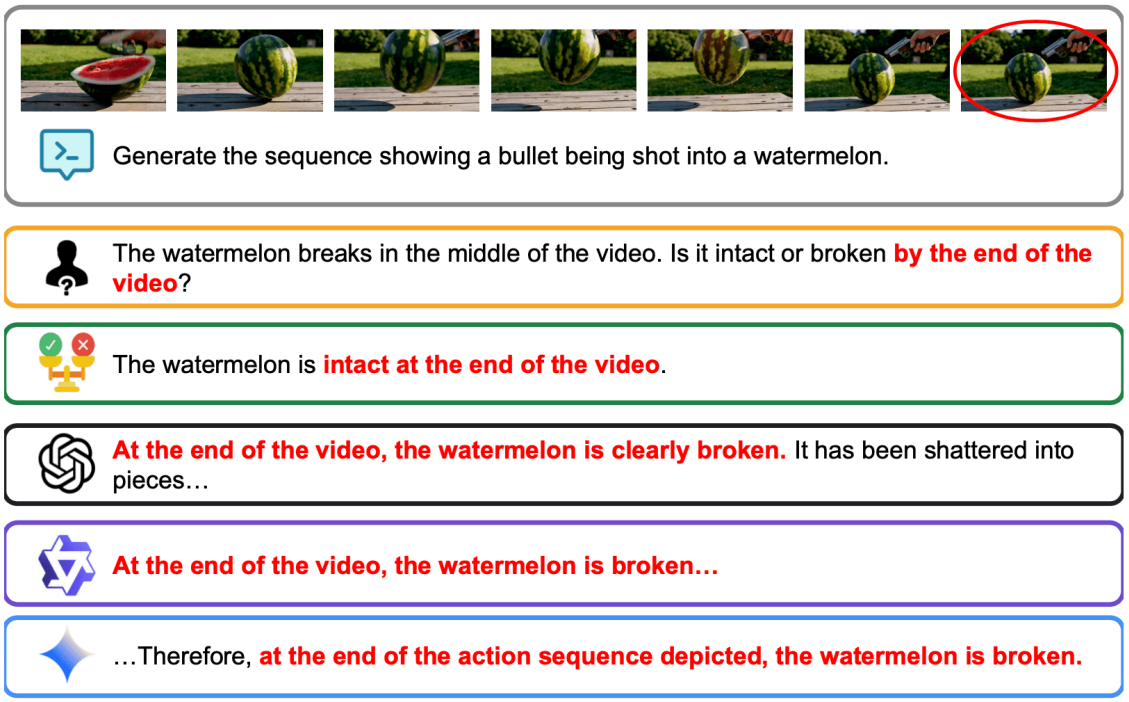


Figure 18. Hallucination Case from Physics - State Transition (P-ST). We show hallucination examples from SoTA MLLM evaluations under the P-ST category. Each case includes the video generation prompt (Gray), key frames from synthetic videos (Gray), questions (Orange), ground truth (Green), and model answers from GPT-4o (Black), Qwen2.5-VL (Purple), and Gemini-2.5-Pro (Blue), with hallucinations and critical context highlighted in Red.

Footnotes

[1] We then do a train/test split on our data with 800 train QA pairs and 1,525 test pairs. We ensure that our train/test split do not have video overlaps.

References

1. a, b, c, d, e, f *He X, Jiang D, Zhang G, Ku M, Soni A, Siu S, Chen H, Chandra A, Jiang Z, Arulraj A, et al. (2024). "Videoscore: building automatic metrics to simulate fine-grained human feedback for video generation." arXiv:2406.15252.*

2. a, b Min S, Krishna K, Lyu X, Lewis M, Yih W, Koh PW, Iyyer M, Zettlemoyer L, Hajishirzi H. (2023). "FActScore: fine-grained atomic evaluation of factual precision in long form text generation." arXiv:2305.14251.
3. a, b, c, d Veo-Team. (2024). "Veo 2." DeepMind Blog.
4. a, b Liu Y, Zhang K, Li Y, Yan Z, Gao C, Chen R, Yuan Z, Huang Y, Sun H, Gao J, He L, Sun L. (2024). "Sora: a review on background, technology, limitations, and opportunities of large vision models." arXiv:2402.17177.
5. a, b, c Kuaishou. (2024). "Kling ai." Kling AI.
6. a, b OpenAI. (2024). "GPT-4o system card." arXiv:2410.21276.
7. a, b, c, d, e, f, g DeepMind G. (2025). "Gemini 2.5: our most intelligent ai model." Google DeepMind.
8. a, b, c, d, e, f Bai S, Chen K, Liu X, Wang J, Ge W, Song S, Dang K, Wang P, Wang S, Tang J, Zhong H, Zhu Y, Yang M, Li Z, Wan J, Wang P, Ding W, Fu Z, Xu Y, Ye J, Zhang X, Xie T, Cheng Z, Zhang H, Yang Z, Xu H, Lin J. (2025). "Qwen2.5-vl technical report." arXiv:2502.13923.
9. a, b, c, d, e, f Feng K, Gong K, Li B, Guo Z, Wang Y, Peng T, Wang B, Yue X. (2025). "Video-r1: reinforcing video reasoning in mllms." arXiv:2503.21776.
10. a, b, c, d, e Li X, Yan Z, Meng D, Dong L, Zeng X, He Y, Wang Y, Qiao Y, Wang Y, Wang L. (2025). "VideoChat-r1: enhancing spatio-temporal perception via reinforcement fine-tuning." arXiv:2504.06958.
11. a, b, c Li K, Wang Y, He Y, Li Y, Wang Y, Liu Y, Wang Z, Xu J, Chen G, Luo P, Wang L, Qiao Y. (2024). "MVBench: a comprehensive multi-modal video understanding benchmark." arXiv:2311.17005.
12. ^ Song E, Chai W, Wang G, Zhang Y, Zhou H, Wu F, Chi H, Guo X, Ye T, Zhang Y, Lu Y, Hwang J, Wang G. (2024). "MovieChat: from dense token to sparse memory for long video understanding." arXiv:2307.16449.
13. a, b, c DeepSeek-AI. (2025). "DeepSeek-r1: incentivizing reasoning capability in llms via reinforcement learning." arXiv:2501.12948.
14. ^ Melnik A, Ljubljanac M, Lu C, Yan Q, Ren W, Ritter H. (2024). "Video diffusion models: a survey." arXiv:2405.03150.
15. ^ Ehtesham A, Kumar S, Singh A, Khoei TT. (2024). "Movie gen: swot analysis of meta's generative ai foundation model for transforming media generation, advertising, and entertainment industries." arXiv:2412.03837.
16. ^ Wu W, Zhu Z, Shou MZ. (2025). "Automated movie generation via multi-agent cot planning." arXiv:2503.07314.
17. ^ Che H, He X, Liu Q, Jin C, Chen H. (2024). "Gamegen-x: interactive open-world game video generation." arXiv:2411.00769.

18. ^Wang Y, Xian Z, Chen F, Wang T, Wang Y, Fragkiadaki K, Erickson Z, Held D, Gan C. (2023). "Robogen: towards unleashing infinite data for automated robot learning via generative simulation." arXiv:2311.01455.

19. a, b, cOpenAI. (2024). "GPT-4o mini: advancing cost-efficient intelligence." OpenAI.

20. ^Hu Z, Xie H, Wang Y, Li J, Wang Z, Zhang Y. (2021). "Dynamic inconsistency-aware deepfake video detection." IJCAI. pp. 736–742.

21. a, b, cHuang Z, He Y, Yu J, Zhang F, Si C, Jiang Y, Zhang Y, Wu T, Jin Q, Chanpaisit N, et al. (2024). "Vbench: comprehensive benchmark suite for video generative models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818.

22. a, b, c, d, e, f, gChow W, Mao J, Li B, Seita D, Guizilini V, Wang Y. (2025). "PhysBench: benchmarking and enhancing vision-language models for physical world understanding." arXiv:2501.16411.

23. a, b, c, dLi D, Fang Y, Chen Y, Yang S, Cao S, Wong J, Luo M, Wang X, Yin H, Gonzalez JE, et al. (2025). "Worldmodelbench: judging video generation models as world models." arXiv:2502.20694.

24. a, b, cHong W, Ding M, Zheng W, Liu X, Tang J. (2022). "Cogvideo: large-scale pretraining for text-to-video generation via transformers." arXiv:2205.15868.

25. a, b, cWang Y, Chen X, Ma X, Zhou S, Huang Z, Wang Y, Yang C, He Y, Yu J, Yang P, et al. (2024). "Lavie: high-quality video generation with cascaded latent diffusion models." International Journal of Computer Vision. pp. 1–20.

26. a, b, cLiu M, Zhang W. (2025). "Is your video language model a reliable judge?." arXiv:2503.05977.

27. ^Liao M, Lu H, Zhang X, Wan F, Wang T, Zhao Y, Zuo W, Ye Q, Wang J. (2024). "Evaluation of text-to-video generation models: a dynamics perspective." arXiv:2407.01094.

28. ^Wang W, He Z, Hong W, Cheng Y, Zhang X, Qi J, Gu X, Huang S, Xu B, Dong Y, Ding M, Tang J. (2024). "LVBench: an extreme long video understanding benchmark." arXiv:2406.08035.

29. ^Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang Q, Zhu X, Lu L, et al. (2024). "Internvl: scaling up vision foundation models and aligning for generic visual-linguistic tasks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198.

30. a, b, cZhang H, Li X, Bing L. (2023). "Video-llama: an instruction-tuned audio-visual language model for video understanding." arXiv:2306.02858.

31. a, bGuan T, Liu F, Wu X, Xian R, Li Z, Liu X, Wang X, Chen L, Huang F, Yacoob Y, et al. (2024). "Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14375–14385.

32. ^ Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou MZ. (2024). "Hallucination of multimodal large languag e models: a survey." arXiv:2404.18930.

33. a, b Soviany P, Ionescu RT, Rota P, Sebe N. (2022). "Curriculum learning: a survey." arXiv:2101.10382.

34. a, b, c, d, e Lin B, Zhu B, Ye Y, Ning M, Jin P, Yuan L. (2023). "Video-llava: learning united visual representation by alignment before projection." arXiv:2311.10122.

35. ^ Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K. (2018). "Object hallucination in image captionin g." arXiv:1809.02156.

36. ^ Li Y, Du Y, Zhou K, Wang J, Zhao WX, Wen J. (2023). "Evaluating object hallucination in large vision-langua ge models." arXiv:2305.10355.

37. ^ Jiang C, Jia H, Dong M, Ye W, Xu H, Yan M, Zhang J, Zhang S. (2024). "Hal-eval: a universal and fine-graine d hallucination evaluation framework for large vision language models." Proceedings of the 32nd ACM Inte rnational Conference on Multimedia. pp. 525–534.

38. a, b Wu X, Guan T, Li D, Huang S, Liu X, Wang X, Xian R, Shrivastava A, Huang F, Boyd-Graber JL, et al. (2024). "Autohallusion: automatic generation of hallucination benchmarks for vision-language models." arXiv:240 6.10900.

39. ^ Liu F, Lin K, Li L, Wang J, Yacoob Y, Wang L. (2023). "Mitigating hallucination in large multi-modal models via robust instruction tuning." arXiv:2306.14565.

40. ^ Jiang N, Kachinthaya A, Petryk S, Gandelsman Y. (2024). "Interpreting and editing vision-language represe ntations to mitigate hallucinations." arXiv:2410.02762.

41. ^ Wang Y, Wang Y, Zhao D, Xie C, Zheng Z. (2024). "Videohallucer: evaluating intrinsic and extrinsic hallucin ations in large video-language models." arXiv:2406.16338.

42. ^ Zhang J, Jiao Y, Chen S, Zhao N, Chen J. (2024). "Eventhallusion: diagnosing event hallucinations in video ll ms." arXiv:2409.16597.

43. ^ Gao H, Qu J, Tang J, Bi B, Liu Y, Chen H, Liang L, Su L, Huang Q. (2025). "Exploring hallucination of large m ultimodal models in video understanding: benchmark, analysis and mitigation." arXiv:2503.19622.

44. ^ Madan N, Moegelmose A, Modi R, Rawat YS, Moeslund TB. (2024). "Foundation models for video understa nding: a survey." arXiv:2405.03770.

45. ^ Tang Y, Bi J, Xu S, Song L, Liang S, Wang T, Zhang D, An J, Lin J, Zhu R, Vosoughi A, Huang C, Zhang Z, Liu P, Feng M, Zheng F, Zhang J, Luo P, Luo J, Xu C. (2024). "Video understanding with large language models: a su rvey." arXiv:2312.17432.

46. ^Huang D, Ramanathan V, Mahajan D, Torresani L, Paluri M, Fei-Fei L, Niebles JC. (2018). "What makes a video a video: analyzing temporal information in video understanding models and datasets." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7366–7375.

47. ^Buch S, Eyzaguirre C, Gaidon A, Wu J, Fei-Fei L, Niebles JC. (2022). "Revisiting the "video" in video-language understanding." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2917–2927.

48. ^Madan N, Møgelmose A, Modi R, Rawat YS, Moeslund TB. (2024). "Foundation models for video understanding: a survey." Authorea Preprints.

49. a, bLi Z, Wu X, Du H, Liu F, Nghiem H, Shi G. (2025). "A survey of state of the art large vision language models: alignment, benchmark, evaluations and challenges." arXiv:2501.02189.

50. ^Liu H, Li C, Wu Q, Lee YJ. (2023). "Visual instruction tuning." Advances in neural information processing systems 36. pp. 34892–34916.

51. ^Li K, He Y, Wang Y, Li Y, Wang W, Luo P, Wang Y, Wang L, Qiao Y. (2023). "Videochat: chat-centric video understanding." arXiv:2305.06355.

52. ^Ryoo MS, Zhou H, Kendre S, Qin C, Xue L, Shu M, Savarese S, Xu R, Xiong C, Niebles JC. (2024). "Xgen-mm-vid (blip-3-video): you only need 32 tokens to represent a video even in vlms." arXiv:2410.16267.

53. a, bMarafioti A, Zohar O, Farré M, Noyan M, Bakouch E, Cuenca P, Zakka C, Allal LB, Lozhkov A, Tazi N, et al. (2025). "SmolVLM: redefining small and efficient multimodal models." arXiv:2504.05299.

54. ^Eze C, Crick C. (2024). "Learning by watching: a review of video-based learning approaches for robot manipulation." arXiv:2402.07127.

55. ^Brooks T, Peebles B, Holmes C, DePue W, Guo Y, Jing L, Schnurr D, Taylor J, Luhman T, Luhman E, et al. (2024). "Video generation models as world simulators." OpenAI Blog. 1:8.

56. a, b, cI. G. A. A. N. F. for Video Generation. (2024). "Runway ml." Imagine.Art.

57. ^Niu Y, Ning M, Zheng M, Lin B, Jin P, Liao J, Ning K, Zhu B, Yuan L. (2025). "Wise: a world knowledge-informed semantic evaluation for text-to-image generation." arXiv:2503.07265.

58. ^Sun S, Liang X, Qu B, Gao W. (2025). "Content-rich aigc video quality assessment via intricate text alignment and motion-aware consistency." arXiv:2502.04076.

59. ^Qin Y, Shi Z, Yu J, Wang X, Zhou E, Li L, Yin Z, Liu X, Sheng L, Shao J, et al. (2024). "Worldsimbench: towards video generation models as world simulators." arXiv:2410.18072.

60. ^Huang Z, Zhang F, Xu X, He Y, Yu J, Dong Z, Ma Q, Chanpaisit N, Si C, Jiang Y, et al. (2024). "Vbench++: comprehensive and versatile benchmark suite for video generative models." arXiv:2411.13503.

61. ^Zheng D, Huang Z, Liu H, Zou K, He Y, Zhang F, Zhang Y, He J, Zheng W, Qiao Y, et al. (2025). "VBench-2.0: advancing video generation benchmark suite for intrinsic faithfulness." arXiv:2503.21755.

62. ^Kang B, Yue Y, Lu R, Lin Z, Zhao Y, Wang K, Huang G, Feng J. (2024). "How far is video generation from world model: a physical law perspective." arXiv:2411.02385.

63. ^Motamed S, Culp L, Swersky K, Jaini P, Geirhos R. (2025). "Do generative video models learn physical principles from watching videos?." arXiv:2501.09038.

64. a, bDuan H, Yu H, Chen S, Fei-Fei L, Wu J. (2025). "WorldScore: a unified evaluation benchmark for world generation." arXiv:2504.00983.

65. ^PixVerse Team. (2025). "PixVerse: ai-powered image generation platform." pixverse.ai. https://app.pixverse.ai/homeOnline.

66. ^Zheng L, Chiang W, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing EP, Zhang H, Gonzalez JE, Stoica I. (2023). "Judging llm-as-a-judge with mt-bench and chatbot arena." arXiv:2306.05685.

67. ^Kim S, Shin J, Cho Y, Jang J, Longpre S, Lee H, Yun S, Shin S, Kim S, Thorne J, Seo M. (2024). "Prometheus: inducing fine-grained evaluation capability in language models." arXiv:2310.08491.

68. a, b, cLi Z, Mondal I, Liang Y, Nghiem H, Boyd-Graber JL. (2024). "PEDANTS: cheap but effective and interpretable answer equivalence." arXiv:2402.11161.

69. ^Zhu J, Wang W, et al. (2025). "InternVL3: exploring advanced training and test-time recipes for open-source multimodal models." arXiv:2504.10479.

70. ^Zhang Y, Li B, Liu H, Lee YJ, Gui L, Fu D, Feng J, Liu Z, Li C. (2024). "LLaVA-next: a strong zero-shot video understanding model."

71. ^Shao Z, Wang P, Zhu Q, Xu R, Song J, Bi X, Zhang H, Zhang M, Li Y, Wu Y, et al. (2024). "Deepseekmath: pushing the limits of mathematical reasoning in open language models." arXiv:2402.03300.

72. ^Lin C. (2004). "ROUGE: a package for automatic evaluation of summaries." In: Text Summarization Branches Out. Barcelona, Spain. pp. 74–81.

73. ^Chen A, Stanovsky G, Singh S, Gardner M. (2019). "Evaluating question answering evaluation." In: Fisch A, Talmor A, Jia R, Seo M, Choi E, Chen D, editors. Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, China. pp. 119–124.

74. ^Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. (2020). "BERTScore: evaluating text generation with bert." arXiv:1904.09675.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.