VidHalluDoctor: Learning Video Differences to Mitigate Hallucinations in Vision-Language Models

Ziyun Dai, Yang Fu, Henghui Ding*
Institute of Big Data, College of Computer Science and Artificial Intelligence,
Fudan University, China
https://github.com/FudanCVL/VidHalluDoctor

Abstract

Large Vision-Language Models (LVLMs) show strong video understanding, yet hallucinations persist, with outputs that misalign with the video content. This stems from two primary causes. First, consecutive frames are often semantically similar, so the model can miss visual details. Second, LVLMs may rely on language priors and pay less attention to visual evidence. To mitigate video hallucinations, we propose **VidHalluDoctor**, a training framework that enhances fine-grained discrimination and reduces reliance on language bias. VidHalluDoctor builds video difference samples to improve visual detail recognition, counterfactual samples to reduce prior bias, and consistency-aware reinforcement learning with designed rewards that evaluate both reasoning quality and consistency between the reasoning and answer. To better evaluate video hallucinations, we introduce **VidHalluEva**, a comprehensive benchmark comprising 10,000 videos that covers three progressive hallucination types across diverse scenarios and multi-view settings. Extensive experiments demonstrate that VidHalluDoctor significantly reduces hallucinations while maintaining strong video understanding. Code and dataset are released.

1 Introduction

Recent advances in Large Language Models (LLMs) [5, 33] have promoted the development of Large Vision-Language Models (LVLMs) [18, 26, 22], which show strong performance on tasks such as image captioning [18] and visual question answering [25, 35]. Despite this progress, imperfect vision–text alignment often causes LVLMs to produce hallucinations [45, 20, 24], in which the model's output is inconsistent with the visual content. Such hallucinations not only undermine model reliability, but also significantly constrain practical deployment in the real world.

Incorporating video modality into LLMs further increases hallucination challenges for LVLMs. Unlike static visual hallucinations [36, 43, 16] in image understanding, videos consist of temporally ordered frames with evolving content [7]. Consequently, LVLMs must not only comprehend individual frame content but also understand events over time, infer their temporal relationships, and remain robust to viewpoint changes introduced by camera motion or multi-view capture.

To characterize video hallucinations in LVLMs, we introduce a progressive taxonomy based on video hallucination content, as shown in Figure 1(a): (1) Object Hallucination: errors in static object existence and attributes (e.g., color, location); (2) Event Hallucination: misidentification of dynamic events and interactions; (3) Temporal Hallucination: incorrect event sequencing. This hierarchical taxonomy comprehensively encompasses current video hallucinations, where continuous objects compose events, and these events are further organized in temporal order.

*Corresponding author (henghui.ding@gmail.com).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

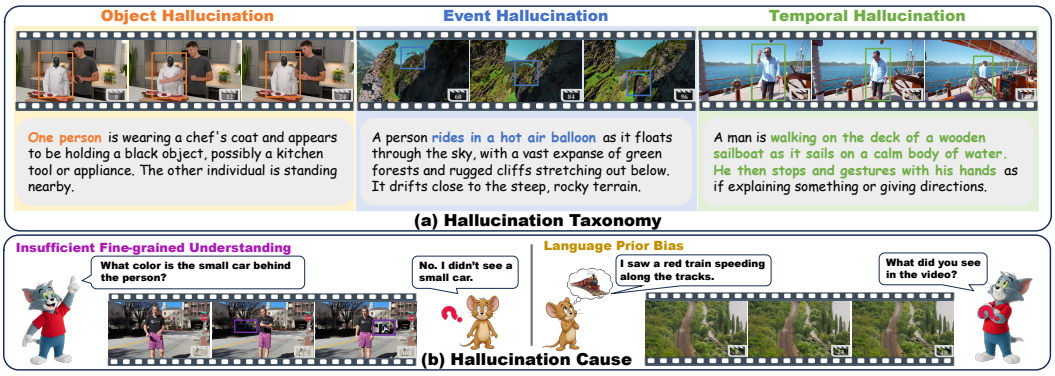

Figure 1: Overview of video hallucination taxonomy and causes in LVLMs. **(a) Progressive three types video hallucination**, spanning object-level attribute errors, event-level action misidentification, and temporal-level event sequence errors. **(b) Hallucination causes**, including insufficient fine-grained understanding and language prior bias from statistical co-occurrence patterns in training data.

Video hallucinations in LVLMs stem from two primary sources, as shown in Figure 1(b). First, consecutive video frames often exhibit high semantic similarity despite visual differences. Insufficient fine-grained understanding causes models to overlook visual changes, resulting in detail loss and hallucinations. Second, language prior bias causes LVLMs to associate objects that frequently co-occur in training corpora while neglecting actual visual content. For example, LLMs trained to associate trains with railways may hallucinate train presence in videos showing only tracks. Existing video hallucination mitigation methods primarily adopt Direct Preference Optimization (DPO) [11, 8, 13] and improved decoding strategies [43, 2], which inadequately enhance fine-grained understanding or address language prior bias through targeted training.

To strengthen fine-grained understanding capabilities, mitigate language prior bias, and enhance reasoning abilities in LVLMs, we propose **VidHalluDoctor**, a novel **Vid**eo **Hallu**cination **Doctor** training framework that integrates specialized data strategies with a progressive training mechanism. VidHalluDoctor builds specialized training data comprising video difference samples that feature semantically similar but visually distinct videos with variations in objects, events, and temporal sequences, as well as counterfactual samples presenting uncommonly associated objects and events. The training mechanism employs a two-stage post-training. First, supervised fine-tuning (SFT) with chain-of-thought (CoT) data establishes foundational reasoning capabilities. Second, reinforcement learning (RL) with CoT data, which we refer to as **consistency-aware RL**, further improves performance through a reward model that evaluates format compliance, answer correctness, reasoning quality, and the consistency between reasoning and answers.

To better evaluate video hallucinations, we introduce **VidHalluEva**, a **Vid**eo **Hallu**cination **Eva**luation benchmark comprising 10,000 videos and 10,000 questions. VidHalluEva covers different hallucination types (object, event, and temporal) and supports three question formats: binary-choice, multiple-choice, and open-ended. To further assess whether models maintain stable understanding across viewpoints, VidHalluEva includes multi-view videos capturing identical events from four different camera views, enabling evaluation of cross-view consistency.

Our main contributions are summarized as follows:

- We propose **VidHalluDoctor**, a novel training framework that mitigates video hallucinations through video difference and counterfactual data strategy, combined with consistency-aware RL to enhance fine-grained understanding and reduce language prior bias.
- We present **VidHalluEva**, a comprehensive benchmark comprising 10,000 videos and 10,000 questions, covering object, event, and temporal types video hallucination, and evaluates models across diverse scenario settings including popular, random, and adversarial scenarios. A multi-view consistency assessment is proposed for the first time, providing a systematic evaluation.
- Extensive experiments demonstrate that **VidHalluDoctor** achieves state-of-the-art performance, significantly reducing hallucinations while enhancing video understanding capability.

2 Related Work

Large Vision-Language Model. Leveraging the remarkable understanding capabilities of LLMs and integrating multimodal encoders, LVLMs have achieved progress in both image-text understanding [18, 25, 35] and video comprehension [15, 19, 42, 17]. While early LVLMs primarily focused on image-text tasks, recent research has extended LVLMs to video modality by building upon image-based foundations to address temporal and sequential frame processing. Video-ChatGPT [29] employs temporal aggregation techniques to integrate cross-frame information, Video-LLaVA [22] uses temporally-sensitive attention to manage inter-frame dependencies, and LLaVA-NeXT-Video [17] introduces linear scaling to process longer videos, capturing extended temporal dependencies.

Video Hallucination Mitigation. Existing mitigation approaches can be broadly categorized into two main strategies: preference-based training and improved decoding methods. Preference-based methods have gained prominence through Direct Preference Optimization, which aligns model outputs with human or AI-provided preferences. PaMi-VDPO [8] introduces an online preference learning framework generating negative samples through video augmentation, and VistaDPO [13] proposes hierarchical DPO optimizing text-video preferences at multiple levels. Decoding-based approaches include temporal contrastive decoding [43] to reduce event-related hallucinations, and MASH-VLM [2] disentangling spatial and temporal representations. However, existing methods often fail to sufficiently enhance the model's ability to focus on fine-grained details or mitigate language priors. To bridge this gap, we propose VidHalluDoctor, which leverages video difference and counterfactual data strategies, combined with consistency-aware RL to enhance fine-grained understanding and reduce language prior bias.

Video Hallucination Benchmark. The development of LVLMs has made video hallucination evaluation increasingly important. VideoHallucer [36] categorizes hallucinations into intrinsic and extrinsic hallucination types, while VidHalluc [16] examines action, temporal sequence, and scene transition hallucinations. EventHallusion [43] evaluates event hallucination. Recent benchmarks have expanded in scope and complexity, VideoHallu [21] uses synthetic videos to test violations of common sense and physical laws, HAVEN [11] presents 6k questions across three dimensions of hallucination causes, aspects, and question formats, and UNSCENE [2] diagnoses action-scene hallucinations using videos with unusual contexts. As shown in Table 1, existing benchmarks often lack a comprehensive taxonomy and rarely evaluate multi-view consistency. We introduce VidHalluEva, establishing a progressive video hallucination taxonomy across 10k videos spanning three hallucination types, while incorporating multi-view evaluation to assess robustness.

3 VidHalluDoctor

In this section, we introduce **VidHalluDoctor**, a novel framework to mitigate video hallucinations in LVLMs. As shown in Figure 2, the framework integrates data strategy and a dual-stage training mechanism to address fine-grained perception errors and reduce language priors. Section 3.1 describes the data strategy regarding video difference and counterfactual samples, while Section 3.2 details the training mechanism comprising SFT and consistency-aware RL.

3.1 Data Strategy

Specialized training data is developed to target the root causes of video hallucinations from three aspects. Video difference samples are designed to sharpen fine-grained visual discrimination. Counterfactual examples are included to diminish reliance on language priors. CoT annotations are utilized to strengthen logical reasoning.

Video Difference Samples. Consecutive video frames often appear semantically identical despite containing subtle visual differences. Insufficient fine-grained perception causes models to overlook these changes, resulting in detail loss and hallucinations. To resolve this issue, we construct video difference samples consist of video pairs that remain semantically similar but visually distinct. This forces the model to detect subtle variations and achieve precise video comprehension. Details are provided in Appendix Section A.3.

As shown in Figure 2(a), video difference samples cover three types based on hallucination taxonomy:

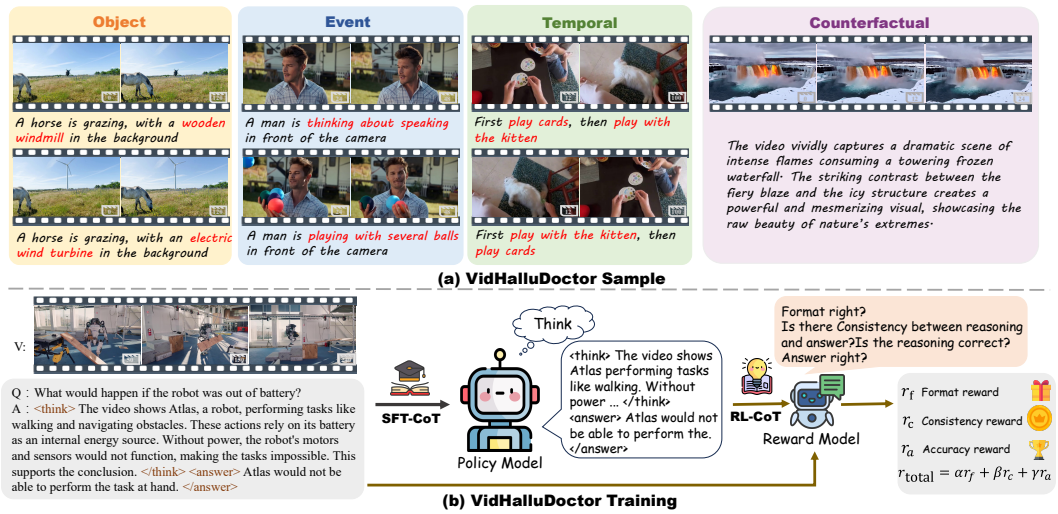

Figure 2: Overview of the VidHalluDoctor framework. **(a) VidHalluDoctor Sample** A specialized data strategy featuring four types of training samples: object differences, event differences, and temporal differences, alongside Counterfactual samples. Differences in video difference pairs are highlighted in red. **(b) VidHalluDoctor Training** A two-stage training mechanism with SFT-CoT to establish reasoning foundations, followed by Consistency-Aware RL.

- **Object difference.** Object attributes change within an identical scene. For example, a wooden windmill is replaced with an electric windmill in the same environment. The videos have similar semantic content but differ in object properties such as material, color, and appearance.
- **Event difference.** The actions or activities performed differ while objects and environments stay similar. For example, a person playing with balls versus the same person standing and thinking. The videos show the same subject in the same setting but performing different actions.
- **Temporal difference.** The same events occur but in reversed order. For example, playing cards then playing with kitten versus playing with kitten then playing cards. The videos contain identical events and objects but differ in event temporal sequence.

Counterfactual Samples. Language prior bias causes models to associate objects that frequently co-occur in training corpora while neglecting actual visual content, which lead to hallucination. For instance, a model might incorrectly describe a train when only tracks in the video. To mitigate this dependence on linguistic priors, counterfactual videos are introduced. As shown in the right of Figure 2(a), these samples feature atypical combinations of objects, environments, and actions that rarely occur in typical scenarios. Examples include dinosaurs playing soccer in the desert, trains flying in the sky. Training on counterfactual data forces models to prioritize visual grounding over linguistic expectations, effectively reducing reliance on language priors.

Chain-of-Thought Annotation. Recent research demonstrates that CoT reasoning enhances LVLMs' analytical capabilities, enabling step-by-step inference grounded in video content. We employ QwenVL-Thinking [3] to generate CoT annotations for all training samples. Given videos, questions, and ground truth answers, the model produces reasoning traces that logically derive correct answers from visual evidence. The prompting strategy is detailed in Appendix A.1.

3.2 VidHalluDoctor Training

As shown in Figure 2(b), the VidHalluDoctor training mechanism consists of two stages: SFT to establish foundational reasoning capabilities with CoT data, followed by consistency-aware RL that leverages video difference pairs to enhance fine-grained discrimination while penalizing hallucinations in reasoning processes.

Supervised Fine-Tuning. In the first stage, we perform SFT on the constructed dataset comprising counterfactual samples, VQA data, and their CoT annotations. The training objective minimizes the negative log-likelihood of generating correct responses conditioned on video and question inputs:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(v,q,a)\sim\mathcal{D}} \left[\log P_\theta(a|v,q)\right], \tag{1}$$

where v represents the video input, q denotes the question, a is the target response including CoT reasoning and the final answer, and θ represents model parameters. This stage enables the model to learn structured reasoning patterns.

Consistency-Aware Reinforcement Learning. In the second stage, we use Group Relative Policy Optimization (GRPO) [31] to further improve model performance through a designed reward mechanism. The reward function has three components: format reward r_f for response structure, consistency reward r_c for reasoning alignment, and correctness reward r_a for answer accuracy.

Consistency Reward Design. The consistency reward r_c evaluates two aspects: (1) whether the reasoning process contains hallucination, and (2) whether the reasoning aligns with the final answer. We use DeepSeek-Chat [23] as the auxiliary LLM judge. Given the policy model's CoT reasoning CoT_θ, the ground truth reasoning CoT_{gt}, the model's answer a_θ, and the ground truth answer a_{gt}, the judge evaluates four criteria: correctness of the final answer, logical coherence of the reasoning, factual accuracy, and consistency between reasoning and answer. The judge outputs a single integer score from 1 to 10, which is normalized to obtain r_c:

$$r_c = \frac{\text{LLM}_{\text{judge}}(\text{CoT}_\theta, \text{CoT}_{\text{gt}}, a_\theta, a_{\text{gt}})}{10}, \tag{2}$$

yielding $r_c \in [0.1, 1.0]$. Higher scores indicate that the model's reasoning is accurate, well-grounded, and logically leads to the correct answer, while lower scores penalize hallucinated reasoning or inconsistency between the reasoning trace and the final answer. The detailed evaluation prompt and scoring rubric are provided in Appendix A.2.

Asymmetric Correctness Reward. For video difference pairs (v_1, v_2) with the same answer choices but different correct answers (a_1, a_2), we assign asymmetric rewards:

$$r_a(v_i, a) = \begin{cases} +1 & \text{if } a = a_i \text{ (correct)} \\ -1 & \text{if } a \neq a_i \text{ (incorrect).} \end{cases} \tag{3}$$

This creates contrastive learning between video pairs. For example, when the model selects option A for v_1 where A is correct, it receives reward $+1$, but selecting the same option A for v_2 where B is correct yields reward -1. This asymmetric structure forces the model to recognize that the same answer choice can be correct for one video but incorrect for its visually similar counterpart.

Total Reward and Optimization. The total reward combines all components:

$$r_{\text{total}} = \alpha r_f + \beta r_c + \gamma r_a, \tag{4}$$

where α, β, and γ are weights balancing different rewards. With the default setting α=0.1, β=0.4, γ=0.5, the reward components operate on comparable scales: $\beta r_c \in [0.04, 0.4]$ and $\gamma r_a \in \{-0.5, +0.5\}$, ensuring that both consistency and correctness contribute meaningfully to the total reward. Hyperparameter sensitivity analysis is provided in Appendix A.6. We adopt GRPO, which samples a group of G outputs $\{o_1, o_2, \ldots, o_G\}$ for each prompt and computes advantages by normalizing rewards within the group. The GRPO objective is:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{(v,q)\sim\mathcal{D}} \frac{1}{G} \sum_{i=1}^{G} \Big[\min\left(\rho_i \hat{A}_i,\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta_{\text{kl}} D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Big], \tag{5}$$

where $\rho_i = \frac{\pi_\theta(o_i|v,q)}{\pi_{\theta_{\text{old}}}(o_i|v,q)}$ is the importance sampling ratio with $\pi_{\theta_{\text{old}}}$ being the behavior policy used for rollout, $\hat{A}_i = \frac{r_{\text{total},i} - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}$ is the group-normalized advantage, ϵ is the clipping parameter, and β_{kl} controls the KL penalty strength.

Gradient Analysis for Video Difference Pairs. To provide intuition for how asymmetric rewards encourage fine-grained discrimination, we analyze the within-group gradient for a single prompt. Consider the prompt for video v_1 with correct answer a_1. Among the G sampled outputs, suppose output o_i correctly selects a_1 (receiving $r_a = +1$) while output o_j incorrectly selects a_2 (receiving $r_a = -1$). Since advantages are normalized within this group, the reward gap of 2γ between o_i and o_j ensures that the group-normalized advantages satisfy $\hat{A}_i > 0$ and $\hat{A}_j < 0$, producing the gradient:

$$\nabla_\theta \mathcal{L}_{\text{GRPO}}^{v_1} \propto \hat{A}_i \nabla_\theta \log \pi_\theta(a_1|v_1) + \hat{A}_j \nabla_\theta \log \pi_\theta(a_2|v_1), \tag{6}$$

which increases $\pi_\theta(a_1|v_1)$ and decreases $\pi_\theta(a_2|v_1)$. Symmetrically, for paired video v_2 where a_2 is correct, the same mechanism increases $\pi_\theta(a_2|v_2)$ and decreases $\pi_\theta(a_1|v_2)$. The combined effect drives the model to assign different answers to visually similar videos based on actual content, directly addressing both fine-grained perception errors and language prior bias.

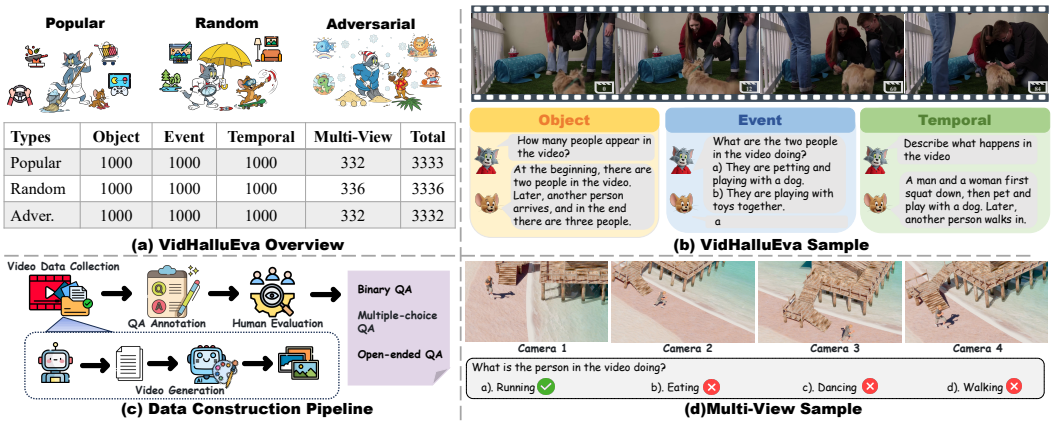

Figure 3: Overview of the VidHalluEva Benchmark. **(a) Data Statistics** VidHalluEva consists of approximately 10,000 video-QA pairs balanced across three scenario settings and three hallucination types alongside multi-view consistency. **(b) Dataset Samples** Representative examples of QA pairs targeting object, event, and temporal. **(c) Data Construction Pipeline** A pipeline comprising diverse video data collection, QA annotation, and human evaluation. **(d) Multi-View Sample** The same event and environment rendered from four distinct camera angles to evaluate whether models maintain consistent and accurate perception across different spatial positions.

4 VidHalluEva Benchmark

In this section, **VidHalluEva** is introduced as a comprehensive benchmark for evaluating video hallucinations in LVLMs. As statistical overview in Figure 3(a), VidHalluEva comprises 10,000 distinct videos paired with 10,000 Question-Answer (QA) pairs, evenly distributed across three hallucination types and three scenarios. The Section 4.1 describes the multi-dimensional design of the benchmark, including hallucination taxonomy, scenario settings, and multi-view consistency. While Section 4.2 details the data construction pipeline and the evaluation protocols.

4.1 Benchmark Design

As shown in Table 1, existing video hallucination benchmarks focus on limited hallucination categories and do not provide assessment of multi-view consistency. To provide comprehensive assessment, VidHalluDoctor is designed with consideration of the following three dimensions. First, a hierarchical video hallucination taxonomy is defined to represent hallucinations ranging from static objects to temporal sequences. Second, three scenario settings are established to evaluate model capabilities across various scenarios. Finally, Multi-View consistency is introduced to evaluate whether model perception remains stable across different camera views.

Hallucination Taxonomy. As shown in Figure 1, We categorize video hallucinations into three types. This hierarchical taxonomy progressively covers video content hallucinations, where static objects form the base, objects and their actions create events, and these events are further organized in temporal order. Based on this taxonomy, we construct the QA pairs in VidHalluEva to specifically target these hallucinations, with the samples displayed in Figure 3(b).

- **Object Hallucination.** Focuses on mistakes related to objects. This includes hallucinating non-existent objects, misidentifying object categories, or incorrectly attributing object properties such as color, shape, size, and spatial location. For example,model might claim a person appears in a video when only a robot is present.
- **Event Hallucination.** Concerns errors in recognizing actions, interactions, and activities in videos. This includes inventing non-existent events, misidentifying actions, or incorrectly describing interactions between objects. For example, model might report that a person rides in a hot air balloon when actually paragliding.
- **Temporal Hallucination.** Deals with the order of events wrong. This refers to the LVLM misrepresenting the temporal order of events. For example, model incorrectly describes a man walking on the deck before gesturing with his hands when the actual sequence shows him stopping to gesture first and then walking.

Table 1: Comparison of VidHalluEva with existing video hallucination benchmarks. VidHalluEva is the largest in scale and covers the most comprehensive set of hallucination and question types. Besides, VidHalluEva provides multi-view consistency evaluation.

Benchmark	Questions	Videos	Question Types			Hallucination Types			Multi-View
			Binary	Multi-Choice	Open	Object	Event	Temporal	
VideoHallucer [36]	1,800	948	✓	✗	✗	✓	✗	✓	✗
VideoHallu [21]	3,233	987	✗	✗	✓	✓	✓	✓	✗
EventHallusion [43]	711	400	✓	✗	✓	✗	✓	✗	✗
VidHalluc [16]	9,295	5,002	✓	✓	✓	✗	✓	✓	✗
HAVEN [11]	6,497	3507	✓	✓	✓	✓	✓	✓	✗
UNSCENE [2]	4,078	1,320	✓	✗	✓	✗	✓	✗	✗
VidHalluEva (Ours)	**10,000**	**10,000**	✓	✓	✓	✓	✓	✓	✓

Scenario Settings. We define three scenarios to evaluate video hallucinations across diverse contexts.

- **Popular.** Videos focus on common daily-life scenarios, such as driving, household chores, gaming, and shopping, including videos captured from first-person perspectives. These represent the typical real-world activities and tasks.
- **Random.** Videos sampled from a wide range of sources to ensure extensive coverage of diverse environments and content types. It serves as a general assessment to evaluate model performance across various real-world contexts.
- **Adversarial.** Comprising counterfactual videos that combine objects, environments, and actions in unlikely ways. By including unusual cases such as camels on grasslands or refrigerators in forests, we evaluate whether models perceive the visual content rather than being misled by language priors.

Multi-View Consistency. Recognizing that viewpoint of camera influences model's perception, we use Blender (a 3D graphics engine) to render multi-view videos of identical event from diverse views, as shown in Figure 3(d). This helps us to evaluate the multi-View consistency, determining whether model maintains stable and accurate responses when observing the same event from different views.

4.2 Data Construction Pipeline

We propose a data construction pipeline, as shown in Figure 3(c), to build high-quality evaluation samples. This pipeline encompasses video data collection from natural datasets and generative data, the QA annotation combines LLM with human verification, and a standardized evaluation protocol designed to quantify model's performance.

Video Data Collection. Video data is sourced from eight datasets: Activitynet-QA [41], Perception test [30], EgoTextVQA [44], MMWorld [12], NExT-QA [38], MSR-VTT [39], UltraVideo [40] and LLaVA-Video-178K [17]. Videos are classified into the three scenario settings (popular, random, adversarial). Due to the lack of counterfactual videos in natural datasets, we use the WAN model [34] to generate videos for the adversarial setting. Five annotators filter the videos to ensure quality.

Question-Answer Annotation. Following the video collection, we prepare the QA pairs for videos. For videos with pre-existing QA pairs, the original pairs are directly used. For the remaining cases, we utilize LLM to formulate questions grounded in video captions, with the prompt provided in Appendix B.1. Qwen-VL [35] is employed to provide initial answers. Finally, all QA pairs are audited and manually corrected by ten human annotators to ensure factual alignment with the videos. For every hallucination type, the questions are structured into three formats: binary, multiple-choice, and open-ended, maintained at a distribution ratio of 3:3:1.

Evaluation Protocol. To provide a comprehensive assessment of model performance, three evaluation metrics are implemented according to question formats and dimensions. Accuracy serves as the primary metric for binary and multiple-choice questions, representing the percentage of correct answers. For open-ended questions, we use GPT-4o [1] to rate responses on a 4-level scale (0–3 points). Furthermore, to evaluate the robustness of perception, Multi-View consistency is measured by analyzing the consistency of multiple-choice responses across four distinct viewpoints of the same event. Scoring rubrics and the evaluation prompts are detailed in Appendix B.2.

Table 2: Results on **VidHalluEva** benchmark. OE: Open-Ended score (0-3). Cons.: Multi-View Consistency (0-2). Acc: Combined accuracy of Binary and Multiple-choice questions. For each column, the best average result is in **bold** and the second best is underlined.

		Object			Event			Temporal			Multiview		Overall	
Model	**Setting**	Binary	Multi.	OE	Binary	Multi.	OE	Binary	Multi.	OE	Multi.	Cons.	Acc	OE
LLaVA-NeXT-Video	Popular	61.46	42.40	0.42	64.45	61.03	0.82	47.11	52.46	1.18	89.16	1.63	58.46	0.81
	Random	77.30	79.87	1.38	76.02	74.52	1.45	47.97	51.39	1.15	78.57	1.45	68.99	1.33
	Adver	90.79	68.95	1.18	78.16	69.16	1.24	47.85	48.71	0.82	75.60	1.27	68.17	1.08
	Average	76.52	63.74	0.99	72.88	68.24	1.17	47.64	50.86	1.05	81.11	1.45	65.21	1.07
LLaVA-OneVision	Popular	67.24	59.53	0.58	71.52	75.16	1.18	62.31	59.96	1.53	93.67	1.77	68.89	1.10
	Random	85.87	91.43	1.68	85.01	88.01	1.62	63.81	66.17	1.59	89.88	1.65	81.10	1.63
	Adver	93.79	85.44	1.21	87.15	87.58	1.61	67.60	67.60	1.21	87.65	1.54	82.18	1.34
	Average	**82.30**	78.80	1.16	**81.23**	83.58	**1.47**	64.57	64.57	1.44	90.40	1.66	77.39	1.36
VideoChat2	Popular	45.40	34.05	0.55	67.67	59.53	1.24	50.75	38.97	1.62	84.04	1.51	53.06	1.14
	Random	68.31	80.51	1.86	72.16	73.88	1.53	51.82	41.33	1.55	77.98	1.35	66.09	1.65
	Adver	88.01	74.09	1.29	81.80	76.23	1.30	59.01	44.85	1.15	81.02	1.31	71.78	1.25
	Average	67.24	62.88	1.23	73.88	69.88	1.36	53.86	41.72	1.44	81.01	1.39	63.64	1.34
Qwen2.5-VL	Popular	64.45	58.03	0.58	69.16	81.37	1.21	64.45	83.73	1.82	93.07	1.70	72.62	1.20
	Random	88.44	90.36	1.94	86.30	90.79	1.68	67.45	83.30	1.95	89.58	1.60	84.99	1.86
	Adver	82.44	87.58	1.32	80.09	87.15	1.47	65.88	78.97	1.13	92.77	1.66	81.67	1.30
	Average	78.44	78.66	1.28	78.52	86.44	1.45	65.93	**82.00**	1.64	91.81	1.65	79.76	1.46
Video-R1	Popular	69.16	63.17	0.52	68.74	76.87	0.86	66.60	76.23	1.80	91.27	1.72	72.37	1.06
	Random	88.01	90.79	1.68	84.80	89.51	1.65	67.88	76.23	2.00	89.29	1.67	83.56	1.78
	Adver	80.30	85.87	1.15	78.80	89.51	1.48	68.24	79.40	0.99	92.47	1.69	81.64	1.21
	Average	79.16	79.94	1.12	77.44	85.30	1.33	67.57	77.29	1.60	91.01	1.69	79.19	1.35
VidHalluDoctor	Popular	61.46	64.67	0.85	74.09	83.30	1.08	69.16	80.09	2.00	94.28	1.80	74.47	1.31
	Random	87.15	91.22	1.65	83.51	87.79	1.68	73.23	80.94	2.06	96.43	1.81	85.31	1.80
	Adver	91.22	89.29	1.48	83.73	92.29	1.44	64.81	73.61	1.25	96.39	1.89	83.97	1.39
	Average	79.94	**81.73**	**1.33**	80.44	**87.79**	1.40	**69.07**	78.21	**1.77**	**95.70**	**1.83**	**81.25**	**1.50**

5 Experiments

Training Data. We construct the VidHalluDoctor dataset for SFT-CoT and RL-CoT training, comprising 8,141 videos and 10,092 questions across binary, multiple-choice, and open-ended formats. Video data is sourced from public datasets including ViDiC [37], EventBench [27], UltraVideo [40] and counterfactual videos generated using the WAN model [34]. We employ Qwen3-VL-Thinking [3] to generate CoT reasoning for all samples. Five annotators manually verify the reasoning traces and answers to ensure quality. This process yields 4,749 SFT-CoT samples and 5,343 RL-CoT samples. Additional construction details are provided in the Appendix A.

Implementation Details. The base model of VidHalluDoctor is Qwen2.5-VL (7B) [4], and all models evaluated on VidHalluEva are at the 7B parameter scale. For the SFT stage, full-parameter fine-tuning is performed with a learning rate of 1×10^{-5} and three training epochs. For the RL stage, we adopt the Easy-R1 [32] framework with full-parameter training. The learning rate of 1×10^{-6} with cosine scheduling and 1 training epoch. We generate three responses per prompt for RL training.

5.1 Experimental Results

Results on VidHalluEva. We evaluate six state-of-the-art LVLMs on VidHalluEva, including LLaVA-NeXT-Video [17], LLaVA-OneVision [15], VideoChat2 [19], Qwen2.5-VL [4], Video-R1 [9], and our VidHalluDoctor. Table 2 presents the comparison across hallucination types, scenario settings, and multi-view consistency. VidHalluDoctor achieves the best overall performance with 81.25% average accuracy and 1.50 open-ended score, outperforming Video-R1 by 0.15 and Qwen2.5-VL by 0.04 on open-ended evaluation. The improvements are particularly significant in adversarial scenarios, with VidHalluDoctor achieving 83.97% accuracy, demonstrating effective reduction of reliance on language prior bias. VidHalluDoctor also achieves the highest multi-view consistency score of 1.83 and 95.70% on multiple-choice questions, substantially outperforming the second-best Qwen2.5-VL at 1.65 and 91.81% respectively, indicating stable understanding across different camera perspectives.

Results on POPE. To assess object-level hallucination, we evaluate VidHalluDoctor on POPE [20], a widely adopted benchmark that probes whether LVLMs hallucinate non-existent objects through polling-based questioning. As shown in Table 3, VidHalluDoctor achieves an accuracy of 87.30% and an F1 score of 85.94%, slightly surpassing the baseline Qwen2.5-VL at 87.28% and 85.83%, respectively. These results indicate that VidHalluDoctor preserves precise object perception capabilities while effectively mitigating broader video hallucinations.

Table 3: Performance on video understanding benchmarks. The best results are shown in bold.

Model	TempCompass				VideoMME				POPE	
	Multi-Choice	Yes-No	Captioning	Caption-Match	Long	Medium	Short	All	Acc	F1
Qwen2.5-VL	**68.80**	71.50	56.70	75.10	48.89	53.56	63.89	55.44	87.28	85.83
VidHalluDoctor	68.18	**73.60**	**59.50**	**75.46**	**49.11**	**55.00**	**70.89**	**58.33**	**87.30**	**85.94**

Results on EventHallusion. To evaluate the effectiveness of VidHalluDoctor against existing hallucination mitigation methods, we compare with three recent methods: PaMi-VDPO [8], DNA-Train [14], and VistaDPO [13]. As shown in Table 4, VidHalluDoctor achieves 68.20%, outperforming all compared methods and the baseline, surpassing VistaDPO by 6.00% and DNA-Train by 6.90%. Although PaMi-VDPO achieves competitive results, VidHalluDoctor still shows improvement.

Table 4: Comparison of methods on EventHallusion.

Method	EventHallusion
PaMi-VDPO	67.90
DNA-Train	61.30
VistaDPO	62.20
Qwen2.5-VL	64.80
VidHalluDoctor	**68.20**

Results on VideoMME. Beyond hallucination evaluation, evaluating on VideoMME [10], a comprehensive benchmark for general video understanding that spans varying durations. As shown in Table 3, VidHalluDoctor consistently outperforms the baseline across all duration settings, improving overall accuracy from 55.44% to 58.33%, with the most notable gain of 7.00% on short videos.

Results on TempCompass. We also assess VidHalluDoctor on TempCompass [28], a general benchmark that evaluates video understanding across diverse task formats. As shown in Table 3, VidHalluDoctor achieves an overall score of 69.19%, outperforming the baseline of 68.03%, with substantial gains on Yes-No (+2.10%) and Captioning (+2.80%) tasks. These results demonstrate that mitigating hallucinations through VidHalluDoctor does not compromise general video understanding capabilities, but rather simultaneously improves them.

5.2 Ablation Study

Table 5 reports the ablation study of each training stage and reward component. SFT-CoT yields improvements on EventHallusion (64.80%→65.77%) but degrades TempCompass from 68.03% to 65.38%, revealing that traditional SFT can overfit hallucination patterns and compromise general video understanding. The RL-CoT stage addresses this effectively, with VidHalluEva accuracy reaching 81.25%, multi-view consistency achieving 1.83, and TempCompass recovering to 69.19%. Removing r_c leads to drops in VidHalluEva (76.37%) and EventHallusion (60.88%), indicating reasoning quality supervision is essential. Removing r_a causes larger declines in VidHalluEva (73.73%) and TempCompass (60.38%), confirming contrastive learning is critical. The full model achieves the best performance, showing r_c and r_a are complementary. Replacing training data with Video-R1 [9] data of equal size yields lower performance, particularly on EventHallusion (62.10% vs 68.20%) and consistency (1.74 vs 1.83), validating the effectiveness of our data strategy.

Table 5: Ablation study on the effectiveness of each training stage and reward component. OE: Open-Ended score (0-3). Cons.: Multi-View Consistency (0-2). r_c: consistency reward. r_a: asymmetric correctness reward. †: trained with Video-R1 data of equal size. The best results are shown in bold.

Models	VidHalluEva			EventHallusion	TempCompass
	Acc	OE	Cons.	Acc	Acc
Qwen2.5-VL	79.76	1.46	1.65	64.80	68.03
+ SFT-CoT	79.69	1.37	1.54	65.77	65.38
+ SFT-CoT & RL-CoT (w/o r_c)	76.37	1.45	1.80	60.88	63.45
+ SFT-CoT & RL-CoT (w/o r_a)	73.73	1.47	1.80	62.59	60.38
+ SFT-CoT & RL-CoT†	81.13	1.47	1.74	62.10	66.90
+ SFT-CoT & RL-CoT (Ours)	**81.25**	**1.50**	**1.83**	**68.20**	**69.19**

6 Conclusion

This paper mitigates and evaluates video hallucinations in LVLMs through two contributions. We propose VidHalluDoctor, a training framework combining video difference, counterfactual data strategies, and consistency-aware RL to enhance fine-grained discrimination and reduce language prior bias. We introduce VidHalluEva, a comprehensive benchmark with 10,000 videos covering object, event, and temporal hallucinations across diverse scenarios with multi-view consistency evaluation. Extensive experiments demonstrate that VidHalluDoctor achieves state-of-the-art performance.

Acknowledgement

Henghui Ding is in part supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123502). This work was supported by the Science and Technology Commission of Shanghai Municipality under Grant No. 25511103600.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13744–13753, 2025.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, and Keqin Chenand others. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M Asano. Lost in time: A new temporal benchmark for videollms. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025. URL `https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_857/paper.pdf`.

[7] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025.

[8] Xinpeng Ding, Kui Zhang, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. *arXiv preprint arXiv:2504.05810*, 2025.

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

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.

[11] 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.

[12] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024.

[13] Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. *arXiv preprint arXiv:2504.13122*, 2025.

[14] Zhe Huang, Hao Wen, Aiming Hao, Bingze Song, Meiqi Wu, Jiahong Wu, Xiangxiang Chu, Sheng Lu, and Haoqian Wang. Taming hallucinations: Boosting mllms' video understanding via counterfactual video generation. *arXiv preprint arXiv:2512.24271*, 2025.

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[16] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13723–13733, 2025.

[17] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

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

[19] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[20] 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.

[21] Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Fuxiao Liu, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*, 2025.

[22] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984, 2024.

[23] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[24] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

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

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[27] Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, and Xiangyang Ji. Eventbench: Towards comprehensive benchmarking of event-based mllms. *arXiv preprint arXiv:2511.18448*, 2025.

[28] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

[29] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024.

[30] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.

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

[32] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[34] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[36] 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.

[37] Jiangtao Wu, Shihao Li, Zhaozhou Bian, Jialu Chen, Runzhe Wen, An Ping, Yiwen He, Jiakai Wang, Yuanxing Zhang, and Jiaheng Liu. Vidic: Video difference captioning. *arXiv preprint arXiv:2512.03405*, 2025.

[38] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[39] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[40] Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

[41] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019.

[42] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[43] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.

[44] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3363–3373, 2025.

[45] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

A VidHalluDoctor

A.1 Chain-of-Thought Annotation Prompt

To generate chain-of-thought reasoning for all training samples, we employ Qwen3-VL-Thinking with the prompt shown in Figure 5. The prompt instructs the model to analyze the video, formulate logical reasoning connecting visual evidence to the answer, and structure the output in two parts: a detailed thinking process enclosed in $< think >$ tags and the final answer in $< answer >$ tags. The model is explicitly instructed to construct coherent reasoning that starts from visual evidence and leads directly to the provided answer, ensuring the reasoning trace grounds predictions in actual video content.

A.2 Consistency Reward Evaluation Prompt

During the RL-CoT stage, the consistency reward r_c is computed by an auxiliary LLM that evaluates the quality of the policy model's reasoning. Figure 6 shows the evaluation prompt used to score the model's chain-of-thought reasoning against ground truth. The LLM judge assesses four criteria: correctness of the final answer, logical coherence of the reasoning process, factual accuracy, and consistency between the reasoning and the answer. Scores range from 1 to 10, which are then normalized to compute r_c^{reason} as defined in Eq. (3).

A.3 Video Difference Sample Construction

Video difference samples are constructed through three distinct approaches corresponding to the three hallucination types: object, event, and temporal differences.

Object Difference Construction For object-level differences, we utilize video pairs from the ViDiC dataset[37]. ViDiC provides paired videos capturing the same scene with controlled variations in object presence and attributes. These variations include adding or removing objects, changing object colors, altering object materials, and modifying object positions. Each video pair maintains identical scene context while introducing specific object-level changes. These naturally occurring object variations provide high-quality training data for fine-grained object discrimination without requiring synthetic generation.

Event Difference Construction Event-level differences are generated by modifying the actions or activities depicted in videos while preserving the visual context. The construction process involves three steps. First, we extract the original video caption describing the events and actions. Second, we use DeepSeek[23] to transform the caption by changing only the event or action while maintaining consistency with the scene and objects. The prompt for caption transformation is shown in Figure 7. The model is instructed to modify events or actions in the caption without altering the scene context or object descriptions. Third, we input the first frame of the original video along with the transformed caption into the WAN video generation model to synthesize a new video depicting the modified event. This approach ensures that event differences are created systematically while maintaining visual consistency in objects and environments. For quality control, generated videos are manually filtered by annotators to ensure the synthesized events match the transformed captions and maintain visual coherence.

Temporal Difference Construction Temporal-level differences are constructed using videos from EventBench[27], which provides temporal annotations marking distinct events and their time boundaries within videos. First, videos with clear temporal event annotations are selected from EventBench. Second, based on the temporal annotations, each video is segmented into discrete event clips corresponding to individual actions or sub-events. Third, the event sequence is reversed by reordering the segmented clips in opposite temporal order. For example, a video showing "playing with kitten then playing cards" is transformed into "playing cards then playing with kitten" by swapping the temporal order of these event segments. This ensures that temporal difference pairs contain identical events and objects but differ only in their chronological sequence, directly testing models' temporal reasoning capabilities without introducing confounding factors from object or event changes.

A.4 Quality Control and Human Verification

All constructed video difference samples undergo quality verification. For object differences from ViDiC, annotators verify that object changes are clearly visible and semantically meaningful. For event differences generated with WAN, five annotators manually filter generated videos to ensure: (1) the synthesized video matches the transformed caption, (2) visual quality is sufficient for evaluation, and (3) object and scene context remain consistent with the first frame. For temporal differences, annotators verify that event segmentation accurately captures distinct actions and that reversed sequences maintain logical visual continuity. Question-answer pairs for all difference types are reviewed by ten annotators to ensure questions clearly target the specific difference type and answers are factually correct based on video content.

A.5 Results on TVBench

Table 6: Results on TVBench. Act.Cnt: Action Count. Act.Loc: Action Localization. Act.Seq: Action Sequence. Ego.Seq: Egocentric Sequence. Mov.Dir: Moving Direction. Obj.Cnt: Object Count. Obj.Shf: Object Shuffle. Scn.Tr: Scene Transition. Unexp: Unexpected Action. The best results are shown in bold.

Model	Act.Cnt	Act.Loc	Act.Seq	Ego.Seq	Mov.Dir	Obj.Cnt	Obj.Shf	Scn.Tr	Unexp	Overall
Qwen2.5-VL	**35.07**	41.88	69.11	56.00	**25.43**	64.19	32.00	**85.95**	50.00	43.37
VidHalluDoctor	34.51	**43.75**	**71.62**	**58.50**	25.00	64.19	**33.78**	83.24	**51.22**	**43.96**

To further evaluate temporal video understanding, we report results on TVBench [6], a benchmark comprising nine task categories that assess fine-grained temporal reasoning capabilities. As shown in Table 6, VidHalluDoctor achieves an overall accuracy of 43.96%, slightly outperforming the baseline Qwen2.5-VL at 43.37%. VidHalluDoctor shows notable improvements on temporal-related tasks such as action sequence (+2.51%), egocentric sequence (+2.50%), and object shuffle (+1.78%), which aligns with our framework's design for enhancing fine-grained temporal discrimination. Performance on other tasks remains comparable, confirming that hallucination mitigation does not degrade general temporal understanding.

A.6 Hyperparameter Sensitivity Analysis

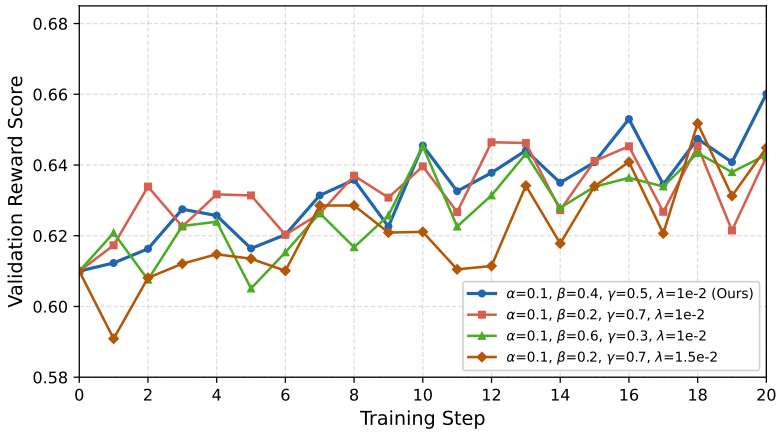

Figure 4: Validation reward score across training steps under different hyperparameter configurations. All configurations show stable upward trends, with the default setting (α=0.1, β=0.4, γ=0.5, λ=1e-2) achieving the highest final reward.

To examine the sensitivity of VidHalluDoctor to reward weight and KL penalty configurations, we conduct experiments by varying β and γ while keeping $\alpha = 0.1$ fixed, as well as varying the KL penalty coefficient λ. Results are summarized in Table 7 and Figure 4.

As shown in Table 7, across all tested configurations, the validation reward score at step 20 remains within a narrow range (0.645–0.660), demonstrating that VidHalluDoctor is robust to hyperparameter

choices. Figure 4 further illustrates the training dynamics, where all configurations exhibit a consistent upward trend in validation reward throughout training. The default setting (α=0.1, β=0.4, γ=0.5, λ=1e-2) achieves the highest final reward of 0.660 and was selected based on validation performance. Notably, increasing λ from 1e-2 to 1.5e-2 slightly reduces the final reward (0.645), suggesting that overly strong KL regularization may constrain policy improvement. Shifting more weight toward the correctness reward (γ=0.7, β=0.2) or the consistency reward (β=0.6, γ=0.3) yields comparable but slightly lower performance, indicating that a balanced weighting between r_c and r_a is beneficial.

Table 7: Hyperparameter sensitivity analysis. All configurations share $\alpha = 0.1$. The default setting is highlighted.

α	β	γ	λ	Step 0 Val Reward	Step 20 Val Reward
0.1	0.4	0.5	1e-2	0.610	**0.660**
0.1	0.2	0.7	1e-2	0.610	0.657
0.1	0.6	0.3	1e-2	0.610	0.655
0.1	0.2	0.7	1.5e-2	0.610	0.645

A.7 Compute Resources

All experiments are conducted on a server equipped with 8× NVIDIA A6000 (48GB) GPUs. The SFT-CoT stage uses 4 GPUs and takes approximately 6 hours to complete. The RL-CoT stage uses 4 GPUs and requires approximately 14 hours. During RL-CoT training, an auxiliary LLM is called via DeepSeek-Chat API to compute the consistency reward r_c, consuming approximately 8.4M tokens in total. The full research project, including preliminary experiments and hyperparameter tuning, required approximately 1.5× the compute reported above.

B VidHalluEva Benchmark

B.1 Data Construction Details

We employ DeepSeek to generate questions for videos without existing QA pairs. Questions are generated based on video captions across three hallucination types (object, event, temporal) and three formats (binary, multiple-choice, open-ended).

Object Hallucination Questions. For object-related questions, the prompts focus on object existence, categories, and properties including color, shape, size, and spatial location. Figure 8 and Figure 9 shows the prompts for binary, multiple-choice, and open-ended formats. The binary format generates Yes/No questions about object presence or attributes. The multiple-choice format creates four options with a specified correct answer, ensuring the other three options are plausible but incorrect. The open-ended format asks descriptive questions about objects in the video.

Event Hallucination Questions. Event-related prompts target actions, interactions, activities, and movement states. The prompts ensure questions directly ask about what happens in the video, testing whether models can accurately recognize dynamic events and interactions.

Temporal Hallucination Questions. Temporal prompts focus on event ordering and temporal relationships. shows the generation prompts. These questions specifically assess whether models can correctly identify the sequence in which events occur.

B.2 Evaluation Protocol Details

Scoring for Multiple-Choice and Binary Questions For binary and multiple-choice questions, accuracy is calculated as the percentage of correct answers. Each question receives 1 point if answered correctly and 0 points otherwise. The final accuracy score is:

$$\text{Accuracy} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Questions}} \times 100\% \tag{7}$$

Scoring for Open-Ended Questions Open-ended questions are evaluated using GPT-4o with the prompt shown in Figure 10. GPT-4o compares the model prediction against the ground truth answer, checks for hallucinations, and provides both a score and brief reasoning. The final open-ended score is the average across all questions. The scoring rubric assigns:

- **0 points**: Completely incorrect or irrelevant response
- **1 point**: Partially correct with severe hallucinations
- **2 points**: Mostly correct with minor hallucinations
- **3 points**: Fully correct without hallucinations

Multi-View Consistency Scoring Multi-view consistency is evaluated exclusively on multiple-choice questions using videos of identical events captured from four different camera viewpoints. The consistency score is calculated based on response agreement across viewpoints:

- **0 points**: Viewpoints yield different answers (inconsistent)
- **1 point**: All four viewpoints yield identical incorrect answers (consistent but inaccurate)
- **2 points**: All four viewpoints yield correct answers (consistent and accurate)

This scoring mechanism evaluates whether models maintain stable predictions across different viewing angles. A score of 2 indicates strong visual grounding with accurate predictions regardless of viewpoint. A score of 1 suggests consistent behavior but incorrect understanding. A score of 0 reveals viewpoint-dependent predictions, indicating reliance on spurious visual patterns rather than robust content understanding.

The final multi-view consistency score is the average across all viewpoint groups, normalized to 0-2 scale. Higher scores indicate better viewpoint invariance and more reliable visual grounding.

B.3 Human Annotation Process

All QA pairs undergo verification by ten trained annotators. Annotators review:

- Question clarity and relevance to the video
- Correctness of ground truth answers
- Plausibility of incorrect options in multiple-choice questions
- Absence of ambiguity or misleading information

Annotators watch each video and verify or correct the generated QA pairs. Any disagreements are resolved through discussion to ensure final consensus. This human verification step ensures high-quality annotations and factual alignment with video content.

B.4 VidHalluEva Samples

To illustrate the diversity and design of VidHalluEva, we present representative examples across three scenario settings and three hallucination types in Figure 11, Figure 12, and Figure 13. These examples demonstrate that VidHalluEva systematically tests video understanding across diverse contexts: Popular scenarios evaluate performance on activities models commonly encounter, Random scenarios assess generalization across varied content, and Adversarial scenarios specifically test whether models truly perceive visual content.

C Limitations

While VidHalluDoctor demonstrates effective hallucination mitigation, several limitations remain. First, the video difference samples rely on video generation models for event-level construction, whose quality may introduce artifacts that affect training. Second, the consistency reward uses an auxiliary LLM as judge, which may itself contain biases that propagate into the reward signal. Finally, VidHalluEva primarily covers object, event, and temporal hallucinations; other hallucination types such as spatial reasoning errors and causal misattribution are not yet addressed.

Symstem message

You are an expert AI assistant specializing in video analysis and reasoning. Your task is to generate a plausible, step-by-step thinking process that logically connects a video to a given question-answer pair.

Prompt

Given a video and a question-answer pair, you need to construct a coherent, logical thought process that starts from the visual evidence in the video and leads directly to the final answer. First, analyze the video and identify key visual elements and events relevant to the question. Then, formulate reasoning that connects observations to the conclusion.Present your output in the following structure. Do not add any text outside of these tags.

Example:

Input:

Video: [A video is provided]

Question: "What is the man doing in the kitchen?“

Answer: "Cooking pasta“

Output:

<think>The video shows a kitchen environment with a man standing near the stove. I can see a pot on the burner with steam rising from it. The man is holding a wooden spoon and stirring something in the pot. On the counter, there are pasta ingredients including a box of spaghetti and tomato sauce. Based on these visual elements, the man is engaged in cooking, specifically preparing pasta.</think><answer>Cooking pasta</answer>

Figure 5: Chain-of-thought annotation prompt for VidHalluDoctor training Prompt for LLM evaluation during RL-CoT training.

Symstem message

Evaluate the consistency and quality of the AI model's reasoning process compared to the ground truth.

Prompt

Ground Truth Information:
<think>{gt_think}</think>
<answer>{gt_answer}</answer>
Model Output to Evaluate:
<think>{model_think}</think>
<answer>{model_answer}</answer>
Evaluation Criteria (Score 1-10):
1. Correctness: Does the model's final answer match the ground truth?
2. Reasoning Logic: Is the thinking process logical, coherent, and step -by-step?
3. Factuality: Does the reasoning rely on correct facts provided in the context?
4. Consistency: Does the thinking process actually lead to the provided answer?
Scoring Guide:
- 1-2: Answer is wrong; reasoning is irrelevant, nonsensical, or completely hallucinated.
- 3-4: Answer is wrong, but reasoning attempts to address the problem; OR answer is correct by pure luck (reasoning is wrong).
- 5-6: Answer is correct, but reasoning has significant logical flaws, contradictions, or skips crucial steps.
- 7-8: Answer is correct; reasoning is mostly correct and logical but may miss minor details or be slightly unclear.
- 9-10: Answer is correct; reasoning is accurate, comprehensive, perfectly aligned with the answer, and demonstrates deep understanding.
Instruction:
Provide ONLY the integer score (e.g., 8) as your response. Do not provide any explanation.

Figure 6: LLM evaluation prompt for computing the consistency reward r_c^{reason} during RL-CoT training. The auxiliary LLM scores the policy model's reasoning against ground truth on a 1–10 scale across four criteria: correctness, reasoning logic, factuality, and consistency.

Symstem message

You are a video event modification expert. Your task is to transform video captions by replacing the MAIN ACTION with a COMPLETELY DIFFERENT, OBVIOUS action

Prompt

Identify the primary action/event in the video (e.g., kicking, running, eating, picking, dusting) and replace it with a distinctly different action that is visually obvious and clear. Keep everything else exactly the same, including all people/subjects (same person, same count, same gender), all objects (same objects, same properties, colors, quantities, sizes), and all locations and scenes. The new action must be dramatically different from the original, not just a slight variation. Generate one transformed caption with the new primary action. Output only the transformed caption text, with no explanations, no JSON, just the modified caption.
Example:
Input:
A man kicking a ball in the stadium.
Output:
A man throwing a ball in the stadium.

Figure 7: Event caption transformation prompt for generating Event Difference samples.

Object

Symstem message
You are watching a video. Here is what happens in the video:
{caption}

Prompt
Generate a binary choice (Yes/No) question about OBJECTS in this video.
Focus on: existence of objects, object categories, object properties (color, shape, size, spatial location).

The question should ask about the video directly, as if the viewer is watching it.
Generate a question where one answer is correct based on what happens in the video.

Output in JSON format:
{{"question": "...", "options": ["Yes", "No"], "answer": "Yes or No"}}

Figure 8: Prompt template for generating Object hallucination questions in VidHalluEva.

Event

Symstem message
You are watching a video. Here is what happens in the video:
{caption}

Prompt
Generate a multiple choice question (4 options) about EVENTS/ACTIONS in this video.
Focus on: actions, interactions, activities, movement states.

The question should ask about the video directly, as if the viewer is watching it.
IMPORTANT: The correct answer MUST be option {correct_option}. Option {correct_option} should be correct based on the video, and the other 3 options should be plausible but incorrect.

Output in JSON format:
{{"question": "...", "options": ["A. ...", "B. ...", "C. ...", "D. ..."], "answer": "{correct_option}"}}

Temporal

Symstem message
You are watching a video. Here is what happens in the video:
{caption}

Prompt
Generate an open-ended question about the TEMPORAL ORDER of events in this video.
Focus on: chronological ordering, sequence of actions, before/after relationships.

The question should ask about the video directly, as if the viewer is watching it.

Output in JSON format:
{{"question": "...", "answer": "..."}}

Figure 9: Prompt templates for generating Event and Temporal hallucination questions in VidHalluEva.

Symstem message

You are an expert evaluator for video question-answering tasks. Your job is to score a model's prediction for an open-ended question.

Prompt

Compare a model's prediction against the ground truth answer, check for hallucinations, assess correctness, and assign a score based on the provided rubric.

Scoring Rubric (0-3 points):

0 points: Completely incorrect. The prediction is entirely wrong or irrelevant to the question.

1 point: Partially correct but with severe hallucinations. The prediction contains some correct information but includes significant false or fabricated details.

2 points: Mostly correct with minor hallucinations. The prediction captures the main content correctly but includes small inaccuracies or minor fabricated details.

3 points: Fully correct with no hallucinations. The prediction accurately answers the question without any fabricated information.

Evaluation Context:

Question: {question}

Ground Truth Answer: {ground_truth}

Model Prediction: {prediction}

Instructions:

1. Compare the model's prediction against the ground truth answer
2. Check for hallucinations (information not supported by the ground truth)
3. Assess correctness (does the prediction answer the question accurately?)
4. Assign a score based on the rubric above

Output Format (JSON only, no markdown):

{{"score": <0-3>, "reason": "Brief explanation for the score"}}

Figure 10: GPT-4o evaluation prompt for open-ended questions in VidHalluEva.

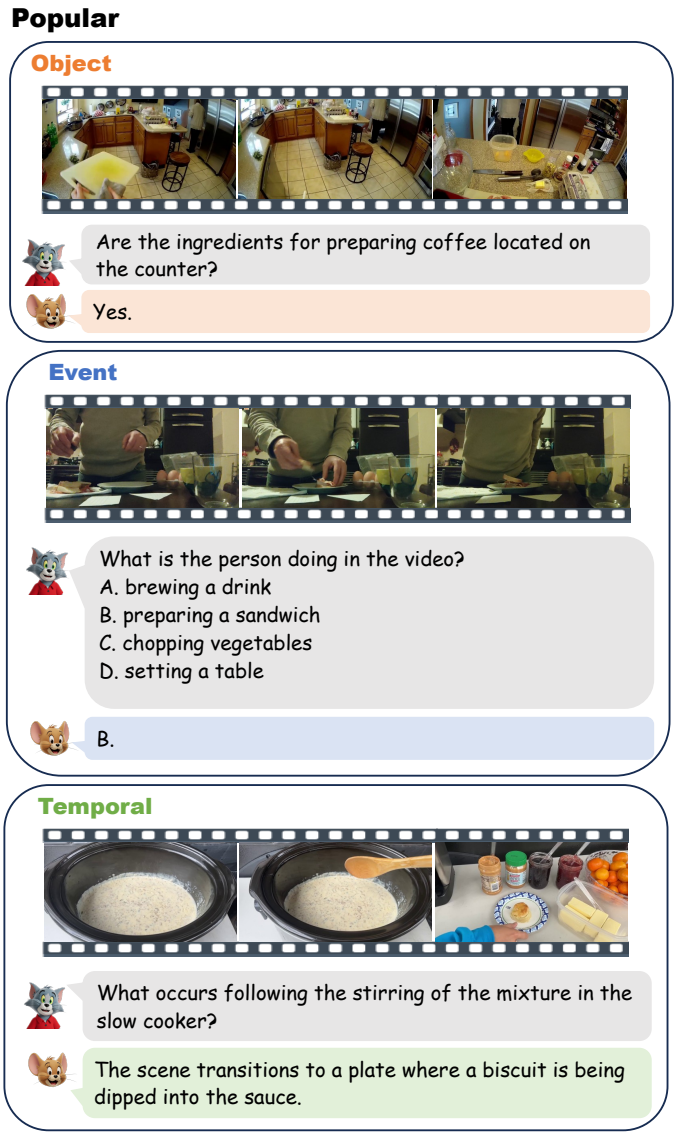

Figure 11: VidHalluEva samples from Popular scenario.

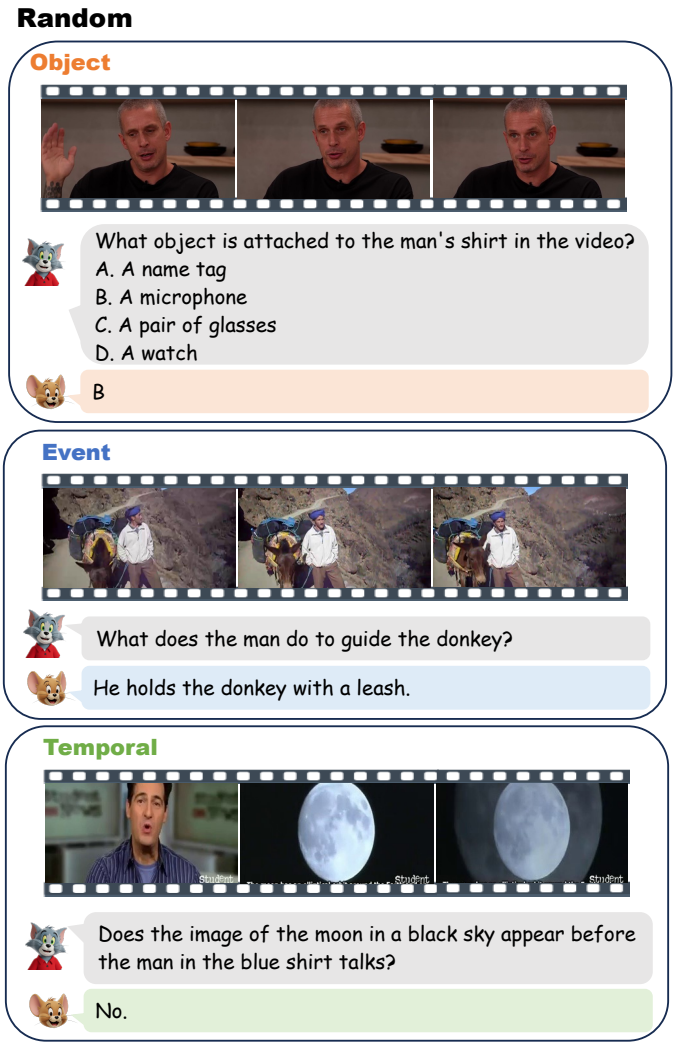

Figure 12: VidHalluEva samples from Random scenario.

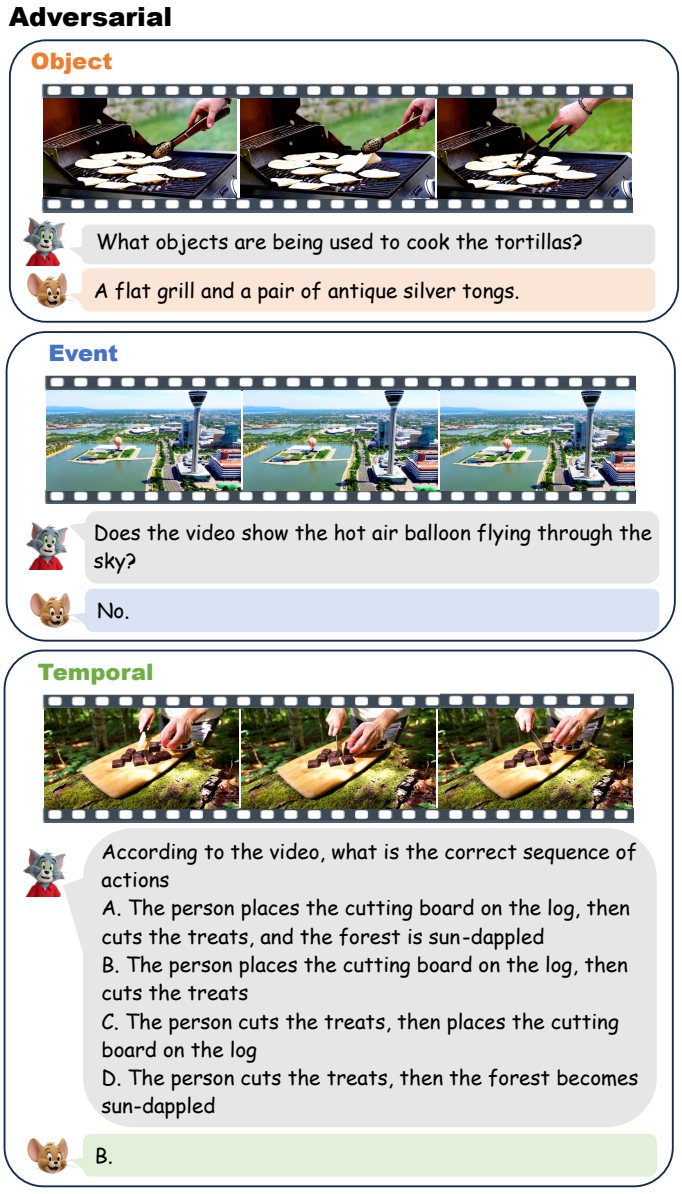

Figure 13: VidHalluEva samples from Adversarial scenario.

D Additional Analysis of Hallucination Sources

Perception and language priors. Insufficient fine-grained perception and excessive reliance on language priors can both contribute to video hallucinations. Perceptual failures can obscure object attributes and action changes across semantically similar frames. Language priors can instead fill in plausible but unsupported content: for example, inferring a train from visible railway tracks, or describing events in a familiar order rather than the order shown. Object hallucinations can arise through either route, from misidentifying a visible object or introducing an absent one. Weak perceptual evidence can also encourage reliance on priors, so the two sources need not act independently.

Difference samples target visually grounded discrimination by varying one factor within a pair, whereas counterfactual samples challenge familiar object–scene–action associations. The two interventions jointly address fine-grained perception and reliance on language priors. Table 8 summarizes the corresponding evaluation axes. Improvements in both popular and adversarial settings show that the model reduces inappropriate reliance on priors while retaining performance on ordinary videos.

Table 8: Complementary evaluation axes. Accuracy is in percent; open-ended (OE) and consistency scores use the original benchmark scales.

Evaluation axis	Qwen2.5-VL	VidHalluDoctor
VidHalluEva Popular accuracy	72.62	74.47
VidHalluEva Popular OE	1.20	1.31
VidHalluEva Adversarial accuracy	81.67	83.97
VidHalluEva multi-view consistency	1.65	1.83
EventHallusion accuracy	64.80	68.20

Event selection. The event substitution prompt in Figure 7 constructs minimal contrast pairs. A replacement action must be visually distinguishable, semantically different, and plausible in the original scene, while preserving the scene, subjects, and objects. Accepted pairs should therefore require recognizing the changed action rather than relying on a changed background or object. Event coverage follows the natural distribution of the real source videos rather than a hand-selected action category list. The same construction and quality-control criteria apply across event types. The gains on independent external benchmarks further demonstrate transfer beyond the constructed pairs.

E Training–Evaluation Overlap Audit

Video-level filtering. De-duplication is performed on the training side before SFT-CoT and RL-CoT. The raw candidate pool shared 459 video basenames with VidHalluEva, all from NExT-QA. A content check compared source-video first-frame CLIP embeddings and manually inspected candidate pairs with similarity above 0.8, identifying 106 additional overlapping videos under different filenames. All 565 overlapping videos were removed before training. The headline results in the main paper were obtained with this filtered training set; these counts describe the candidate pool rather than contamination of the trained model. VidHalluEva was not modified by this filtering.

Table 9: Training-side overlap audit. Content duplicates are additional to basename matches. Zero-overlap entries refer to the retained training data.

Stage	Comparison or removal	Videos
Candidate pool	Basename matches with VidHalluEva	459
Candidate pool	Additional content matches after manual confirmation	106
Before training	Total overlapping videos removed	565
Retained data	SFT/RL training versus VidHalluEva	0
Retained data	Training versus EventHallusion, TempCompass, TVBench, and VideoMME	0

The retained training data was also checked against EventHallusion, TempCompass, TVBench, and VideoMME. The few VideoMME overlaps identified in the candidate pool were removed. A supplementary segment-level audit examined 500 videos sampled from the training-side pool, uniformly sampling five frames per video. Candidate cross-video matches were selected by maximum

frame-pair CLIP similarity using the same 0.8 threshold and manually confirmed. Within this audited subset, the check found no additional overlaps beyond those identified by the first-frame procedure. This segment-level audit extends the content check beyond the first frame.

Question-level audit. We compared 5,074 training questions with the 10,000 VidHalluEva questions after normalization. Six exact shared question templates cover 418 evaluation items (4.18%). Including near-duplicates identified by token Jaccard similarity of at least 0.9 after normalization gives 424 items (4.24%). These are generic formulations such as "What is the man doing in the video?" None of these evaluation items corresponds to the 565 removed overlapping videos. The 5,074 count is the question set used for this text audit, distinct from the 10,092 stage-specific QA training samples described in the main paper.

Benchmark composition and controlled comparison. NExT-QA contributes 1,200 of the 10,000 evaluation items (12%). Newly generated counterfactual and multi-view portions together account for over 40% of the benchmark. These constructed videos and questions reduce direct reuse of public evaluation content. VidHalluDoctor and its Qwen2.5-VL baseline share the same initialization, providing a controlled comparison of the additional training recipe.

F Reliability of the LLM Judges

Different judge roles. The DeepSeek-Chat reward judge evaluates whether the model's reasoning is factually grounded and supports its final answer, using the ground-truth information and the criteria in Appendix A.2. It does not require reproducing one reference reasoning path. In contrast, GPT-4o evaluates open-ended benchmark answers with the 0–3 rubric in Appendix B.2. For benchmark judging, only the final answer extracted from the `<answer>` tag is submitted; the model's full reasoning trace is excluded. This separates benchmark answer evaluation from the reasoning traces used during training.

Table 10: Aggregate judge validation. Human agreement and prompt stability are separate checks with different sampled sets.

Check	Items	Result
GPT-4o versus blind human OE scoring	240	Pearson 0.83; Spearman 0.86; exact agreement 71%; within one point 93%
DeepSeek consistency versus human judgment	160	Agreement on whether reasoning supports the answer: 88%
GPT-4o with a paraphrased prompt	40	Spearman 0.93; exact agreement 82.5%; within one point 100%; mean score difference 0.13
DeepSeek with a paraphrased prompt	40	Spearman 0.70; mean score difference 0.15 on the consistency scoring scale

Disagreement example. Disagreements in the human–GPT-4o check often concern partially correct answers and detail coverage. For a reference describing "a flat grill and antique silver tongs," the prediction "cooked on a grill using tongs" received 3 from the human and 2 from GPT-4o. The answer captures the central action and tools but omits attributes, illustrating different strictness within the same rubric rather than disagreement about the central event.

Aligned versus mismatched reasoning–answer pairs. We additionally evaluated 112 aligned pairs and 48 mismatched pairs, created by combining the same reasoning with a deliberately incorrect answer. On the original 1–10 judge scale, aligned pairs average 9.81 and mismatched pairs average 6.23; the ranking AUC is 0.96. This diagnostic supports the judge's ability to distinguish aligned reasoning–answer pairs from mismatched ones.

Multiple valid reasoning paths. An observed training example yielded three different chains supporting the same correct multiple-choice answer: one used the microphone and performing behavior, another used the interaction between the two people, and the third ruled out distractors using visual evidence. The latter two chains had reported textual similarities of 0.29 and 0.23 to the first. All three received the same total reward of 0.91, demonstrating that different valid reasoning paths can receive the same reward.

G Additional Comparisons and Training Ablations

Same-backbone comparison. Table 11 compares methods using Qwen2.5-VL-7B and the same evaluation script. The comparison includes four decoding-time hallucination mitigation methods and two RL post-training methods. VidHalluDoctor achieves the highest accuracy on both EventHallusion and VidHalluEva among these methods.

Table 11: Additional Qwen2.5-VL-7B comparisons. Both reported metrics are accuracies in percent.

Method	Setting	EventHallusion	VidHalluEva
Qwen2.5-VL	Base	64.80	79.76
+ VCD	Decoding	65.68	80.12
+ ICD	Decoding	66.55	79.93
+ OPERA	Decoding	65.11	80.87
+ TCD	Decoding	67.68	80.02
+ Video-R1	RL	65.53	79.19
+ GRPO	RL	61.12	79.26
VidHalluDoctor	Post-training	**68.20**	**81.25**

Separating data and reward effects. Table 12 varies the training data and reward design under the controlled training recipe. Under standard GRPO rewards, replacing Video-R1 data with VidHalluDoctor data raises EventHallusion accuracy from 61.12 to 65.59. With Video-R1 data fixed, the full reward raises accuracy from 61.12 to 62.10. With VidHalluDoctor data fixed, the full reward raises it from 65.59 to 68.20. The combination performs best. These results demonstrate the complementary contributions of hallucination-oriented data and the consistency-aware reward.

Table 12: Data and reward decomposition on EventHallusion with Qwen2.5-VL-7B. Differences from the base are in percentage points.

Training data	Reward design	Accuracy	Δ versus base
Base model	–	64.80	–
Video-R1	Standard GRPO	61.12	-3.68
Video-R1	VidHalluDoctor	62.10	-2.70
VidHalluDoctor	Standard GRPO	65.59	$+0.79$
VidHalluDoctor	VidHalluDoctor	**68.20**	$+3.40$

The equal-size Video-R1-data variant in Table 5 uses our consistency-aware RL and differs from the Video-R1 baseline in Table 11. It reaches 81.13 on VidHalluEva accuracy, close to 81.25 for the full model, but its EventHallusion accuracy is 62.10 versus 68.20 and its multi-view consistency is 1.74 versus 1.83. Pooled accuracy therefore provides a less discriminative summary of these variants than the complementary hallucination and consistency metrics. With Video-R1 data fixed, the full reward also raises TempCompass from 59.48 for standard GRPO to 66.90.

SFT cold start and stability. SFT-CoT supplies the response format used by the subsequent RL stage. In a format diagnostic at step zero, the raw base model was tested on 100 training prompts with four rollouts per prompt at the training temperature. None of the 400 outputs contained an `<answer>` tag or the required think-then-answer structure. The training prompts in this diagnostic did not explicitly instruct that format. This observation motivates SFT-CoT as the format-learning cold start for RL. The results in Table 5 show that the improvements materialize when SFT-CoT is followed by consistency-aware RL.

Across the four tested reward-weight and KL-penalty configurations in Appendix A.6, the step-20 validation rewards are 0.660, 0.657, 0.655, and 0.645. This narrow range indicates limited sensitivity to these hyperparameter choices within the tested range.

H Generation Quality Control and Data Composition

Automatic and manual checks. Generated samples are screened with CLIP semantic consistency and temporal alignment scores, using the score appropriate to the hallucination type. The documented

decision bands are scores above 0.80 for automatic acceptance, scores below 0.50 for automatic rejection, and the intermediate band for human review. Approximately half of the generated pool enters this review band. Five annotators check visual quality, agreement with the transformed caption, and preservation of scene and objects. These automatic checks complement the human filtering described in Appendix A.3.

Table 13: Failure categories in the generated-video audit. The 38% rejection rate is conditional on a sample of 200 videos from the manual-review pool; category shares are percentages of rejected samples and are rounded.

Failure mode	Share of rejected samples
Action not replaced	38%
Unintended background change	22%
Visible artifacts	23%
Subject drift	17%

The audit sampled 200 videos from the manual-review pool across generation types and found a human rejection rate of 38%. On a double-annotated subset of the audit pool, Cohen's κ was 0.86. QA verification is performed by approximately ten trained annotators as described in Appendix B.3.

Regeneration policy. An initial trial regenerated rejected samples for one additional round after rewriting their prompts. Of these regenerated videos, 87% were rejected again. The final construction policy therefore discards rejected samples instead of repeatedly regenerating them. Every retained sample must satisfy its applicable acceptance checks. This final policy supersedes the initial regeneration procedure and limits repeated generation cost.

Training composition and distribution gap. The training set contains 8,141 videos and 10,092 QA samples, with 4,749 SFT-CoT and 5,343 RL-CoT samples. Rounded RL composition is 62.6% difference samples, 18.7% counterfactual samples, and 18.7% multi-view samples; counterfactual examples account for approximately 20% of SFT. Counts by video path across stages give approximately 53% synthetic or edited videos and 47% real-source videos. These path-based proportions have a different denominator from the RL sample-type proportions.

Real-source videos and regular difference samples keep the training mixture anchored to ordinary video content, while synthesized or edited pairs supply controlled differences that are difficult to obtain naturally. First-frame conditioning during event synthesis and checks for scene, object, and action preservation further constrain unwanted variation. The gains on popular videos and external benchmarks further demonstrate transfer to ordinary video content.

I Temporal Errors and Transition-Artifact Diagnostic

Temporal multiple-choice errors. On the temporal multiple-choice subset of VidHalluEva, accuracy decreases from 82.00 for Qwen2.5-VL to 78.21 for VidHalluDoctor, despite improved temporal open-ended and multi-view results. The per-item analysis found that two-event questions asking which event happens first had approximately twice the error rate of multi-event ordering questions. A recurring failure is treating a scene setup or object state as the first event. In one example, the reference sequence is that pigeons first gather on the ground and then peck at scattered gemstones; the model instead selects an ordering in which scattering the gemstones comes first. This pattern suggests difficulty distinguishing an explicitly observed event from background conditions implied by the scene.

Table 14: Additional external temporal evaluation, in percent. VideoMME columns distinguish temporal perception (TP), temporal reasoning (TR), and overall accuracy.

Model	VideoHallucer	VideoMME		
	Temporal	TP	TR	Overall
Qwen2.5-VL	51.70	67.27	36.72	55.44
Video-R1	13.64	63.64	39.55	54.89
VidHalluDoctor	**55.68**	**69.09**	**41.24**	**58.33**

Transition-artifact check. To examine whether editing cuts drive temporal judgments, we compared two diagnostic conditions using the same temporal questions. Real reordering changes the event order, whereas fake reordering inserts uninformative cuts while keeping the order unchanged. Accuracy is 42% under real reordering and 76% under fake reordering. The 76% accuracy under fake reordering indicates that the model can retain its temporal judgment when cuts are inserted without changing the event sequence, rather than treating cuts alone as evidence of an order change. The consistent gains on independent external temporal benchmarks in Table 14 and TVBench in Appendix A.5 provide further evidence of transfer beyond the constructed reordering samples.

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: Section 1

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Section C

   Guidelines:

   - The answer [N/A] 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: Section 3.2

Guidelines:

- The answer [N/A] 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: Section 5

   Guidelines:

   - The answer [N/A] 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: Partial data is included in the supplemental materials, and the complete data and code will be released upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5, Section A.6

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: Section 5, Section A.6

   Guidelines:

   - The answer [N/A] 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: Section A.7

   Guidelines:

   - The answer [N/A] 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: Our work adheres to the NeurIPS Code of Ethics. We address safety, fairness, and responsible model usage throughout the research.

   Guidelines:

   - The answer [N/A] 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: [Yes]

    Justification: Section 1

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not handle the safety of models, so it is not applicable to our paper.

    Guidelines:

    - The answer [N/A] 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: All datasets, models, and code used in this work are properly licensed.

    Guidelines:

    - The answer [N/A] 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 are in the process of supplying comprehensive documentation, including usage instructions, data schema, and evaluation procedures.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are core components of the method and benchmark pipeline: Qwen3-VL-Thinking generates chain-of-thought annotations; DeepSeek generates questions and transformed captions; DeepSeek-Chat evaluates reasoning–answer consistency during reinforcement learning; Qwen-VL supplies initial answers for benchmark annotation; and GPT-4o scores open-ended benchmark answers. The corresponding prompts and protocols are described in the method and appendix, and additional judge validation is provided in Appendix F.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.