NO PLACE TO HIDE: BENCHMARKING VIDEO HALLUCINATION WITH BACKGROUND-CONTROLLED PAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce VIDPAIR-HALLUC, a new benchmark for evaluating video hallucination in large video models (LVMs) under rigorous and controlled conditions. Unlike previous benchmarks that primarily rely on text-based perturbations or adversarial questions while neglecting the consistency of visual backgrounds, VIDPAIR-HALLUC features video pairs with highly similar backgrounds but distinctly different foreground semantics, enabling precise attribution of model errors to genuine hallucination rather than background variation. The benchmark is constructed through PAIRFLOW, a pipeline that leverages recent advances in text-to-image and video generation to systematically compose stories, generate coherent video clips, and assemble them into adversarial pairs. Covering both spatial and temporal reasoning across ten semantic aspects, VidPair-Halluc comprises 1K high-quality adversarial video pairs and 11K spatio-temporal QA pairs with control over background and foreground variations. We evaluate mainstream LVMs on VidPair-Halluc, and our results show that current models still struggle with robust and fine-grained video understanding in adversarial settings. VidPair-Halluc.

1 INTRODUCTION

Progress in video understanding is driven by large-scale datasets (Fu et al., 2024; Li et al., 2024a; Wang et al., 2024b; Zhou et al., 2024a; Bain et al., 2021; Xiao et al., 2021; Wang et al., 2019) and advanced large video models (LVMs) (Lin et al., 2023a; Bai et al., 2023; Cheng et al., 2024; Li et al., 2023; Maaz et al., 2023; Chen et al., 2024b). A key challenge is video hallucination, where models generate content inconsistent with visual evidence, often due to a capability gap between video encoders and large language models (LLMs) (Huang et al., 2025; Yuan et al., 2024). While LLMs undergo extensive pre-training, weaker video encoders lead to overconfident yet inaccurate outputs (Liang et al., 2024; Cui et al., 2024). To address video hallucination, benchmarks have evolved progressively. Initial works (Yang et al., 2024; Zhang et al., 2024a; Gao et al., 2025) focus on single videos, using spatio-temporal questions to assess basic understanding. Building on this, subsequent studies introduce adversarial textual contexts (see Figure 1 (*b*)), such as misleading questions or distractor answers within single videos (Wang et al., 2024b), to more comprehensively evaluate model robustness to high-level semantic perturbations. Nevertheless, since most LVMs inherently depend on instruction tuning, their predisposition to strictly follow user directives suggests that hallucinations observed in these benchmarks may be confounded by human-induced bias, thereby undermining the objectivity of such evaluations. Recent benchmarks employ paired videos (Guan et al., 2024; Li et al.) to assess fine-grained semantic alignment, yet they mainly rely on text-driven disturbances and CLIP/DINO-based selection without explicitly controlling background consistency. Consequently, hallucination behavior may be confounded by background variation rather than purely reflecting foreground semantic misinterpretations, where the background is query-irrelevant and the foreground query-relevant visual context. While background variation is intrinsic to real-world hallucinations, controlling for similarity serves as a controlled setting that enables more precise attribution to foreground semantics. Such isolation is crucial for disentangling sources of video hallucination, enabling a more systematic understanding of model robustness.

To address these limitations, we propose a hallucination benchmark with background-consistent yet foreground-divergent video pairs, enabling clearer attribution of model errors to hallucination rather than background shifts or semantic misguidance. However, constructing such adversarial pairs in real videos is labor-intensive, as fixed storylines and spatio-temporal dependencies require extensive human intervention. In contrast, image-based VQA has leveraged advances in image inpainting and

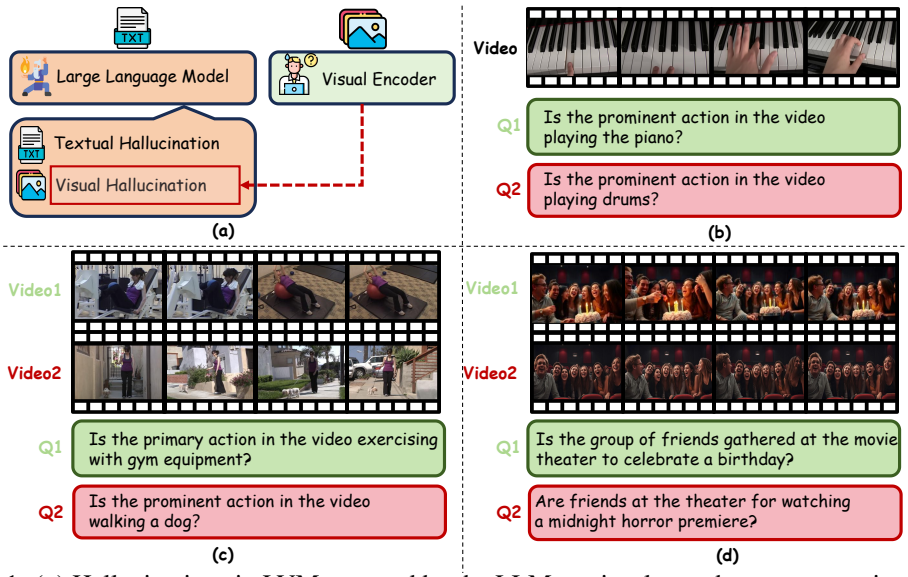

Figure 1: (***a***) Hallucinations in LVMs, caused by the LLM or visual encoder, are categorized based on adversarial sources into textual or visual, where visual hallucination poses a greater challenge. (***b***) Adversarial question benchmarks (*e.g.*, VideoHallucer (Wang et al., 2024b)) mainly induce textual hallucinations by perturbing the LLM. (***c***) VIDHALLUC (Li et al.) introduces video pairs that are visually dissimilar but share similar overall semantics, challenging multimodal understanding. (***d***) Our benchmark constructs video-text pairs with highly similar background but distinctly different foreground semantics, enabling targeted evaluation of hallucinations.

automated pipelines (Zhuang et al., 2024; Zhang et al., 2024b; Wu et al.; 2025c; Liu et al., 2024c; Google DeepMind, 2024), achieving scalable adversarial sample generation. Inspired by this, we look to recent progress in video generation and editing to address scalability in the video domain. Yet, existing approaches remain limited: (***i***) Direct video generation or frame replacement often disrupts background or entity consistency. (***ii***) Video editing lacks stability and fidelity in complex scenes. (***iii***) Generating video pairs from edited key frames and descriptions faces challenges from scarce frame-level captions and high annotation costs.

These challenges motivate us to explore whether advanced text-to-image and video generation models can streamline data collection and construct adversarial video pairs with minimal human intervention. To this end, we propose **PAIRFLOW**, a pipeline of three stages: ❶ ***Story Composition*:** Automatically generate paired stories with controlled dependencies. ❷ ***Video Clip Generation*:** Synthesize coherent clips via advanced text-to-image (T2I) and video generation models. ❸ ***Video Assembly*:** Concatenate clips into adversarial video pairs for robust LVM evaluation. Based on PairFlow, we construct video pairs with similar backgrounds but distinct foregrounds, and introduce **VIDPAIR-HALLUC** (Figure 1 (*d*)) for rigorous hallucination benchmarking. Our main contributions are summarized as follows:

- We propose PAIRFLOW, a data pipeline for constructing background-consistent, foreground-divergent adversarial video pairs, enabling fine-grained evaluation of video hallucination.
- We introduce VIDPAIR-HALLUC, a new benchmark covering both spatial and temporal reasoning from 10 perspectives, featuring 1K high-quality adversarial video pairs and 11K spatio-temporal QA pairs, with rigorous control over background and foreground for precise benchmarking.
- We benchmark mainstream LVMs on VIDPAIR-HALLUC, providing in-depth analysis of their strengths and limitations. We further assess state-of-the-art hallucination mitigation methods, revealing current approaches remain insufficient for robust fine-grained video understanding.

2 VIDPAIR-HALLUC: AN ADVERSARIAL BENCHMARK FOR VIDEO HALLUCINATION

Existing benchmarks typically assess video hallucination either by posing misleading questions on a single video (Wang et al., 2024b; Zhang et al., 2024a; Choong et al.) or by pairing videos with low visual similarity but distinct semantics (Li et al.; Chen et al., 2024c; Guan et al., 2024). A more

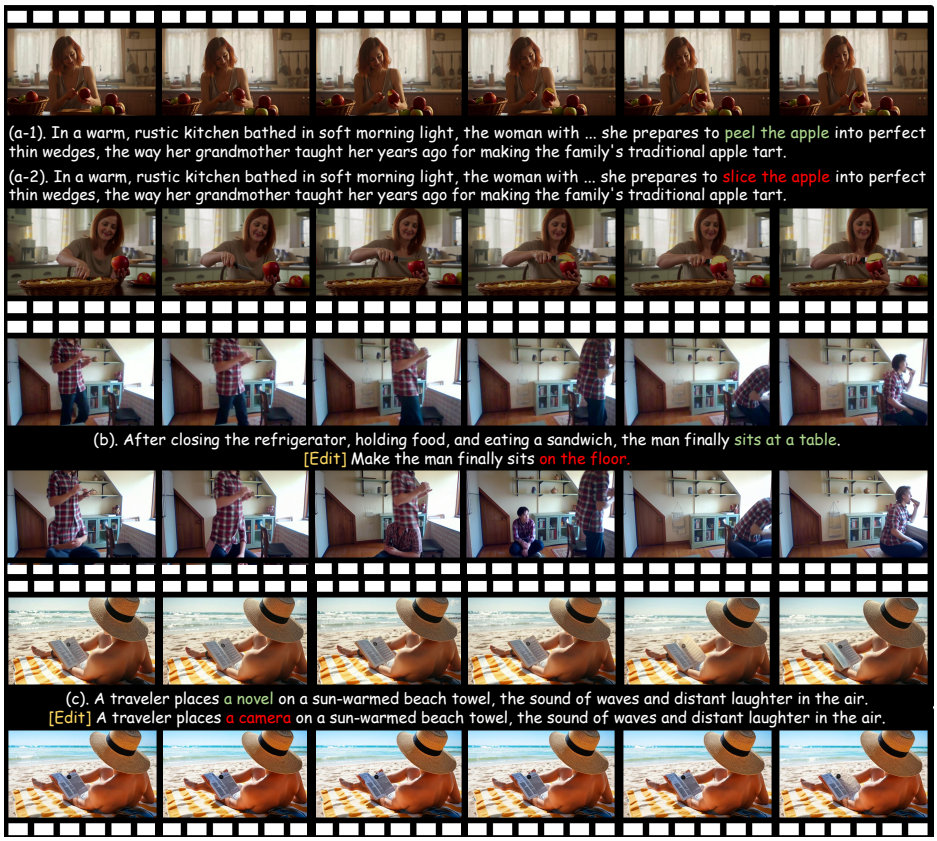

Figure 2: **Challenges in Generating High-Quality Adversarial Video Pairs**. (***a***) Cutting-edge video generation models (*i.e.* HunyuanVideo (Kong et al., 2024)) create pairs with significant semantic differences but fails to maintain consistent visual backgrounds. (***b***) Advanced image editing methods (*i.e.* SeedEdit (Shi et al., 2024)), applied frame-by-frame, lead to noticeable inconsistencies. (***c***) Advanced video-editing techniques (*i.e.* TokenFlow (Geyer et al., 2023)) result in pairs that lack faithfulness, stability and robustness.

demanding and challenging setting involves adversarial pairs that share highly similar backgrounds yet differ in entity semantics, where subtle distinctions are easily missed by LVMs (Guan et al., 2024; Wu et al., 2025b; Chen et al., 2024c). However, constructing such pairs from real videos is costly and impractical, limiting current benchmarks' ability to probe hallucinations under these nuanced conditions. To close this gap, we build on PairFlow and introduce VIDPAIR-HALLUC, which systematically evaluates video hallucination using hierarchical levels of adversarial video pairs.

Benchmark	Number of Question/Videos	Binary QA	Multi-Choice QA	Open-Ended QA	Visual Similarity	Control Pairs	Adversarial
HallusionBench (Guan et al., 2024)	1,129/346	✓	✓	✗	✗	✓	✓
VideoHallucer (Wang et al., 2024b)	1,800/948	✓	✗	✗	✗	✗	✓
Vript-HAL (Yang et al., 2024)	122/122	✗	✓	✗	✓	✗	✗
EventHallusion (Zhang et al., 2024a)	–/400	✓	✗	✓	✗	✗	✗
VIDHALLUC (Li et al.)	9,295/5,002	✓	✓	✓	✗	✓	✓
VIDPAIR-HALLUC (Ours)	11,523/2000	✓	✓	✓	✓	✓	✓

Table 1: Comparison of existing video hallucination benchmarks.

2.1 PAIRFLOW: AN ADVERSARIAL VIDEO PAIRS DATA PIPELINE

Motivations. The fixed storylines and spatio-temporal dependencies in real videos demand extensive human effort, limiting the scalability of adversarial video pair generation with previous methods. In contrast, image-based VQA tasks have benefited from advances in image editing technologies (Zhuang et al., 2024; Zhang et al., 2024b), enabling a shift from hand-crafted corner cases to automated pipelines that efficiently scale up adversarial samples generation (Wu et al.; 2025c; Liu et al., 2024c). Motivated by this progress, we turn to recent advances in video generation and editing as a promising path to overcome the scalability bottleneck in the video domain.

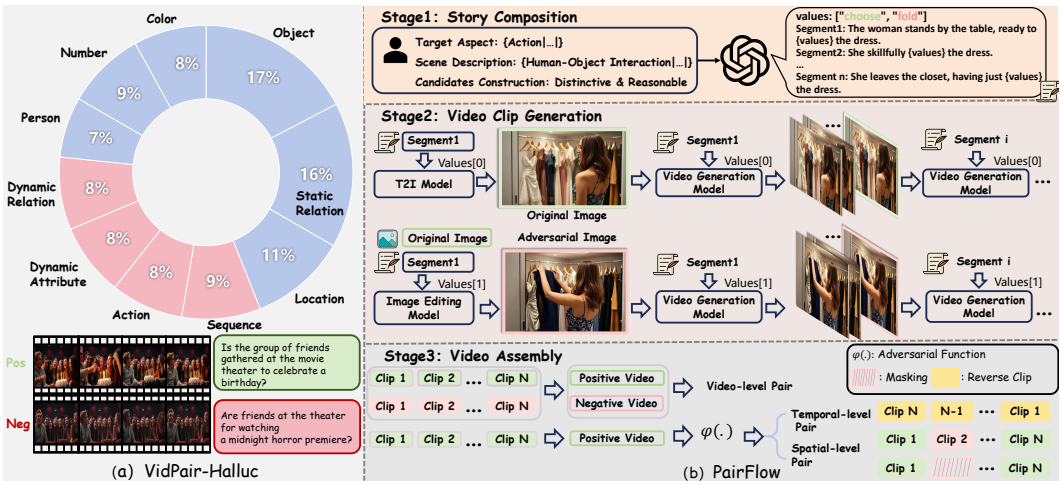

Figure 3: (***a***) VIDPAIR-HALLUC benchmark quantitatively evaluates hallucinations in video models from both temporal (33%) and spatial (67%) perspectives, covering a broad range of semantic aspects. (***b***) The benchmark is constructed via the PAIRFLOW framework, which generates distinctive video-text scenarios, produces video clips with controlled semantic variations, and assembles positive and negative pairs at multiple levels, enabling comprehensive and robust hallucination benchmarking.

Fortunately, advances in video generative technologies indeed offer promising solutions to these challenges. Several intuitive candidate approaches can be envisioned: (***i***) directly generating video pairs (*e.g.* (Bai et al., 2025b)) or editing frame by frame; (***ii***) applying video editing (*e.g.* (Ma et al., 2025; Zhou et al., 2023)); (***iii***) generating video pairs conditioned on edited key frames and descriptions. But these approaches still face several limitations: ❶ Simple video-level generation (*i.e.* Figure 2 (*a*)) or frame-level replacement methods (*i.e.* Figure 2 (*b*)) often fail to maintain entity or background consistency (Geyer et al., 2023), causing unnatural transitions and less plausible adversarial samples . ❷ Current video editing techniques, though promising, still lack the stability and fidelity needed for reliable adversarial editing in complex scenes (*i.e.* Figure 2 (*c*)). ❸ Editing-and-generation strategies are promising, but the lack of frame-level captions and varying frame resolutions in real videos hinder target selection, raising annotation costs and limiting scalability.

Given these observations, we are motivated to explore a more scalable and controllable paradigm. A key question arises: *can advanced text-to-image and video generation models streamline data collection and efficiently construct adversarial video pairs with minimal human intervention?* To address this, we propose **PAIRFLOW**, a novel framework to answer this by reducing manual annotation, while also holding the potential to enhance video-language alignment and video hallucination benchmarking. The details of data construction with PAIRFLOW follows.

Data Construction with PAIRFLOW. PairFlow consists of three stages: story composition, video clip generation, and video assembly. Initially, the process begins with the composition of narratives, where specific elements such as actions and human-object interactions are emphasized. This stage involves constructing story segments using placeholders, allowing for the creation of diverse variations. For instance, a narrative might depict a woman preparing to perform an action like "choose" or "fold" a dress, represented as **{Values}**. Each candidate value is required to be logically consistent with the storyline and significantly different from one another.

In the video clip generation stage, an advanced pretrained text-to-image (T2I) model is employed to produce original images, which serve as the foundation for creating adversarial samples. These images undergo enhancement through sophisticated image editing techniques, ensuring that while the background remains consistent, the foreground exhibits significant distinction for high-quality adversarial samples. A noteworthy aspect of this process is the sequential generation of video segments, denoted as V_i. Each segment V_i is generated using the last frame V_{i-1}^L of the preceding clip V_{i-1} in conjunction with the narrative script corresponding to segment V_i. This approach is mathematically denoted as:

$$V_i = f(V_{i-1}^L, S_i), \quad i \geq 2, \tag{1}$$

where $f(\cdot)$ denotes a video generation model which integrates the last frame V_{i-1}^L and the story script S_i to produce the subsequent video segment V_i, ensuring each clip has a length of L.

Figure 4: Examples from VIDPAIR-HALLUC. Each row shows a positive-negative video-level pair, with four frames per video (left: positive, right: negative). The first three rows illustrate spatial reasoning, while the last two focus on temporal reasoning. Binary QA and MCQ are provided for each pair, highlighting contrasting answers under highly similar visual contexts.

Finally, video assembly seamlessly integrates these clips into coherent sequences by crafting both positive and negative video versions through meticulous temporal and spatial pairings. The adversarial function $\varphi(.)$ is pivotal in this process, employing multiple operations to generate adversarial samples at both temporal and spatial levels. Temporally, each clip is individually reversed and then rearranged in reverse order $[V_N, V_{N-1}, \ldots, V_1]$, resulting in a completely inverted video sequence. Spatially, the process involves randomly replacing segments within video-level pairs and applying masking operations to the modified adversarial videos. Such adversarial samples pose significant challenges to the model's ability to comprehend specific foreground semantics across different clips.

Benchmark Quality Assurance. We adopt a compact, multi-stage pipeline that couples model-assisted curation with human review. Trained raters first vet GPT-4.1 story scripts for coherence and distinct, context-appropriate endings; annotators then verify each generated clip in Label Studio (Tkachenko et al., 2020-2025) for description fidelity and baseline video quality. For adversarial pairs (cf. Fig. 4), reviewers retain only cases with *highly similar backgrounds* and *clear foreground semantic differences*, removing instances with weak or ambiguous foreground cues. Finally, for every validated pair we instantiate exactly one QA using targeted templates spanning ten spatial/temporal hallucination types, and a separate rater group confirms strict question–video alignment. Full protocols, rater instructions, and diagnostics are provided in the Appendix.

2.2 VIDEO-LEVEL ADVERSARIAL PAIRS

At the video level, we construct adversarial pairs by assembling positive and negative videos that differ in overall semantics or storyline as shown in Figure 4. The negative video in each pair is

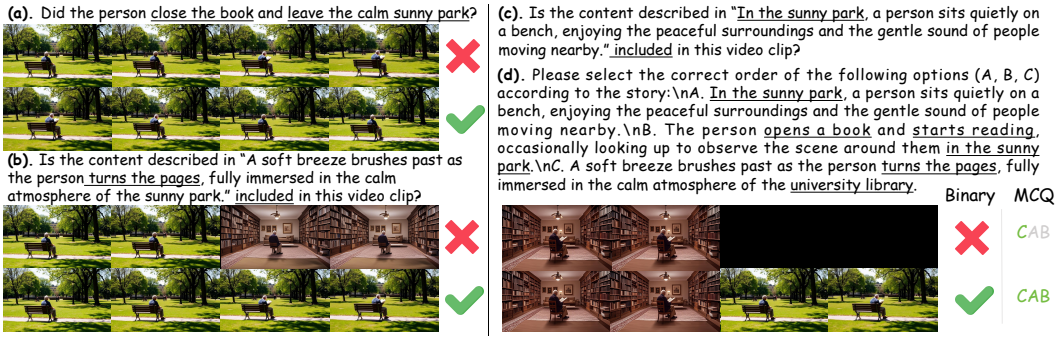

Figure 5: Illustration of the increased challenge posed by temporal-level (***a***) and spatial-level (***b***) adversarial pairs compared to video-level adversarial pairs. Each spatial-level pair is accompanied by both binary QA (***c***) and MCQ (***d***) tasks. Notably, the MCQ format requires assessing and ranking the relevance of clip-level descriptions, thereby demanding a finer-grained understanding of spatial and temporal relationships within the video.

generated by replacing or altering the entire content, ensuring that its global meaning deviates from the positive counterpart. This design aims to test the model's ability to capture semantic consistency at a holistic level. Detecting such discrepancies is fundamental, as a reliable model should distinguish between videos with entirely different narratives, even when superficial visual similarities exist.

2.3 Temporal-Level Adversarial Pairs

Temporal-level adversarial pairs are constructed by perturbing the temporal order of clips within a video. For negative samples, we mainly employ complete sequence reversal, since it preserves temporal semantics to the greatest extent and serves as the most fundamental way to test the model's capability in temporal reasoning and understanding. While other operations like shuffling or segment-level reordering can also disrupt chronological flow, sequence reversal is particularly suited for evaluating temporal logic as shown in Figure 5(*a*).

2.4 Spatial-Level Adversarial Pairs

At the spatial level, we construct adversarial pairs by masking, replacing, or subtly modifying specific clips within a video, which introduces local inconsistencies such as missing objects or altered scenes. For example, changes to the book held by an elderly person or modifications in the surrounding environment, as shown in Figure 5(*b,c*), can disrupt object continuity while preserving the overall narrative structure. The binary QA is specifically designed to assess the model's sensitivity to spatial details and object-level coherence. Building on this, the MCQ format further evaluates the model's ability to maintain temporal consistency, addressing the common challenge where LVMs fail to accurately capture the foreground semantics of individual segments, resulting in contradictions or biases in temporal reasoning. By leveraging our proposed PairFlow framework, we can efficiently generate clip-level adversarial samples for fine-grained evaluation. Compared to clip-level captioning on real-world videos with fixed story logic, our method leverages customized story contexts to construct adversarial pairs with highly similar backgrounds and distinct foreground semantics. This design enables more targeted and lightweight evaluation of both spatial and temporal understanding.

3 Experiment

3.1 Experimental Settings

Models. We evaluate 15 mainstream open-source and close-source LVMs: Video-ChatGPT (Maaz et al., 2023), Video-LLaVA (Lin et al., 2023a), VideoChat2 (Li et al., 2023), Video-LLaMA2 (Cheng et al., 2024), PLLaVA (Xu et al., 2024), Qwen2.5-VL (Bai et al., 2025a), R1-OneVision (Yang et al., 2025), ThinkLite-VL (Wang et al., 2025b), ShareGPT4Video (Chen et al., 2024b), LLaMA-VID (Li et al., 2024b), VILA1.5 (Lin et al., 2023b), Gemini-2.5-Pro, Gemini-2.5-Flash (Team et al., 2024), GPT-4o (Hurst et al., 2024) and GPT-5-mini (OpenAI, 2025).

Evaluations. For fair comparison, all models are evaluated on binary QA, MCQ, and open-ended description tasks. For binary QA, to assess model robustness and performance, particularly in the

Table 2: Performance comparison of open-source and closed-source LVMs on the VIDPAIR-HALLUC for binary, multi-choice, and open-ended QA. All metrics are reported as percentages. Human results serve as the upper bound. "-" denotes "N/A". **Bold** font indicates the best performance, while underlining denotes the second best.

Method	Params	Language Model	Binary			Multi-Choice		Open-Ended
			wAcc ↑	FP (∼ 0)	Pct. Diff (∼ 0)	F1 ↑	vAcc ↑	Desc. ↑
Open-source LVMs								
Video-ChatGPT (Maaz et al., 2023)	7B	LLaMA-7B	24.58	33.66	16.21	6.10	0.0	27.70
Video-LLaVA (Lin et al., 2023a)	7B	Vicuna-7B-v1.5	31.90	35.30	23.26	53.52	24.60	34.74
VideoChat2 (Li et al., 2024a)	7B	Vicuna-7B-v0	18.47	2.12	-45.89	38.03	0.79	42.25
Video-LLaMA2 (Cheng et al., 2024)	7B	LLaMA2-7B	21.48	43.00	33.41	62.91	42.86	40.85
Qwen2.5-VL-Instruct (Bai et al., 2025a)	7B	-	41.66	48.71	14.77	61.50	42.86	45.07
R1-OneVision (Yang et al., 2025)	7B	-	35.52	27.24	2.20	29.23	28.33	42.37
ThinkLite-VL (Wang et al., 2025b)	7B	-	31.86	33.40	4.79	43.29	37.27	48.39
ShareGPT4Video (Chen et al., 2024b)	8B	LLaVA-Next-8B	19.53	**1.67**	-44.95	38.50	0.08	34.27
PLLaVA (Xu et al., 2024)	13B	Vicuna-13B-v1.5	32.06	45.72	-6.74	8.92	2.38	41.31
LLaMA-VID (Li et al., 2024b)	13B	Vicuna-13B-v1.5	18.79	53.13	49.75	33.80	15.08	30.52
VILA1.5 (Lin et al., 2023b)	13B	-	18.44	15.18	-16.01	58.22	37.30	38.03
Closed-source LVMs								
Gemini-2.5-Flash (Team et al., 2024)	-	-	29.83	18.97	-8.82	62.44	39.68	50.23
Gemini-2.5-Pro (Team et al., 2024)	-	-	**49.15**	13.07	**-2.12**	**67.32**	43.36	**54.68**
GPT-4o (Hurst et al., 2024)	-	-	26.97	29.16	10.97	59.15	38.10	47.89
GPT-5-mini (OpenAI, 2025)	-	-	29.33	19.65	3.28	64.82	**45.28**	49.33
Human	-	-	74.32	9.28	4.37	89.21	79.66	-

context of adversarial video pairs and question pairs, we introduce several metrics: **Question Pair Accuracy**, **Video Pair Accuracy** and **Yes Percentage Difference (Pct. Diff)** as follows.

Question Pair Accuracy. This metric evaluates whether a model can consistently answer all instances within an adversarial pair correctly. Specifically, let $\mathbb{V}_i$ denote the set of videos in the i-th adversarial pair, and $\mathbb{Q}$ denote the set of question pairs. The Question Pair Accuracy is defined as:

$$\text{qAcc} = \frac{\sum_{i,k} \mathbb{1}\left(\bigwedge_{V \in \mathbb{V}_i} b_{\mathcal{M}}(V, q(i,k))\right)}{|\mathbb{Q}|}, \quad b_{\mathcal{M}}(V,q) \in \{0,1\} \tag{2}$$

where $b_{\mathcal{M}}(V,q)$ is a binary indicator of correctness for model $\mathcal{M}$ on video V and question q, and $\mathbb{1}(\cdot)$ is the indicator function. This metric highlights the model's ability to provide consistent and robust answers across all elements of an adversarial pair, reflecting its resilience to both visual and textual adversarial perturbations.

Video Pair Accuracy. Symmetrically, let $\mathbb{V} = \bigcup_{i \in \mathcal{I}} \mathcal{V}_i$ be the set of videos across all adversarial pairs. We say a specific video $V \in \mathcal{V}_i$ is answered *consistently* if the model is correct on *all* questions in $\mathbb{Q}$ when evaluated on V. The video-pair accuracy is defined as follows:

$$\text{vAcc} = \frac{1}{|\mathbb{V}|} \sum_{i \in \mathcal{I}} \sum_{V \in \mathcal{V}_i} \mathbb{1}\left(\bigwedge_{q \in \mathbb{Q}} b_{\mathcal{M}}(V,q) \right). \tag{3}$$

To provide an intuitive summary of a model's overall robustness to video hallucination, we further report a weighted average, wAcc, defined as:

$$\text{wAcc} = \frac{|\mathbb{Q}|\,\text{qAcc} + |\mathbb{V}|\,\text{vAcc}}{|\mathbb{Q}| + |\mathbb{V}|}, \tag{4}$$

where $|\mathbb{Q}|$ and $|\mathbb{V}|$ denote the numbers of question pairs and videos, respectively. This aggregates the two accuracies in proportion to their sample sizes.

Yes Percentage Difference. This metric measures the deviation between the proportion of "yes" responses given by the model and that in the ground truth. Formally,

$$d_y = \frac{\sum_{(V,q) \in \mathcal{A}} \left[\mathbb{1}(\mathcal{M}(V,q) = \text{"yes"}) - \mathbb{1}(y(V,q) = \text{"yes"})\right]}{|\mathcal{A}|}, \tag{5}$$

where $\mathcal{A}$ is the set of all (video, question) pairs, $\mathcal{M}(V,q)$ is the model's answer, and $y(V,q)$ is the ground truth. A value of $|d_y|$ close to 1 indicates a strong bias towards a particular answer, while a value near 0 suggests balanced predictions.

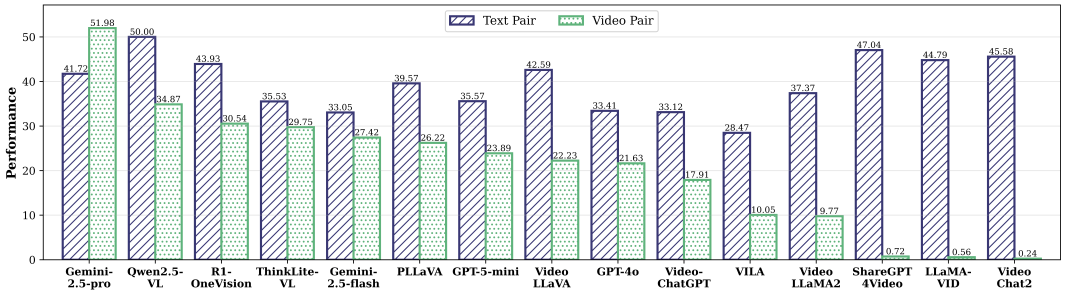

Figure 6: Performance comparison between text-pair and video-pair hallucinations.

For MCQ, besides *qAcc*, we also report the F1 Score, a standard metric for multi-class classification:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{6}$$

where precision and recall are computed from true positives, false positives (FP), and false negatives.

Implementation Details. For story script generation, we utilized the GPT-4.1 (Hurst et al., 2024) API to automatically produce narrative descriptions. Based on these scripts, we employed the advanced open-source text-to-image model FLUX (Labs, 2024) to generate keyframes. Subsequently, precise foreground semantic editing was performed on these keyframes using the advanced image editing model SeedEdit (Shi et al., 2024). Leveraging both the edited images and the corresponding story segments, we further synthesized 480p video clips with the advanced open-source video generation model Wan2.1-14B and Wan2.2-14B (Wang et al., 2025a). All data processing and inference were conducted on a cluster of eight H100 GPUs. For open-ended questions evaluation, we employ GPT-4o (Hurst et al., 2024) and follow the settings in (Zhang et al., 2024a). We recruited three English-proficient evaluators with computer science backgrounds to assess the benchmark results. To reduce bias, question-answer pairs were randomized to avoid consecutive similar types.

3.2 Main Results

As shown in Table 2 and Figure 7 (c), closed-source LVMs lead and are better calibrated. Gemini-2.5-Pro performs best overall, achieving the top Binary wACC and low FP, state-of-the-art Multi-choice results, and the strongest open-ended descriptions. GPT-5-mini attains the highest vAcc. Critically, low FP means fewer spurious claims, which is vital for safety-critical or cautious decision pipelines where errors of commission are costly. High wACC signals broad robustness across adversarial video and text pairs and across clip- and video-level spatial semantics, not just isolated wins. Among open-source models, Video-LLaMA2 excels on Multi-choice (temporal reasoning), whereas Qwen2.5-VL reaches strong wACC but with elevated FP, exposing a calibration gap. Overall, the results highlight a persistent trade-off between spatial sensitivity and hallucination control. Closing this gap will require coordinated progress in temporal reasoning, spatial understanding, and risk-aware calibration.

3.3 Further Analysis

Qualitative Results. Figure 7 (a) visualizes Qwen2.5 VL Instruct 7B embeddings for VidPair-Halluc adversarial pairs. With strict background control, positive and negative samples largely overlap in both video and text, indicating weak separation of relation polarity. The learned representation underweights foreground semantics that are most informative for hallucination detection and instead appears biased toward background cues. This bias likely propagates to cross-modal alignment and blurs the distinction between genuine and adversarial pairs. In summary, the model captures scene context and style more than the subtle object and action evidence required for robust hallucination control.

Synthesis Viability. Figure 7 (b) underscores the feasibility of generative data synthesis. As video generators advance, single-pass human vetting rates rise markedly, indicating higher fidelity and controllability. Complementary results in the appendix show strong performance from prevalent editing models. Together, these trends suggest that next-generation image and video generators will become reliable building blocks for adversarial multi-modal pairs pipelines. As model capability improves, the synthesized samples will gain in realism and diversity, making them increasingly effective for both training (*e.g.* Hierarchical Preference Optimization in (Huang et al., 2025; Fu et al., 2025b)) and hallucination evaluation (Bai et al., 2025b).

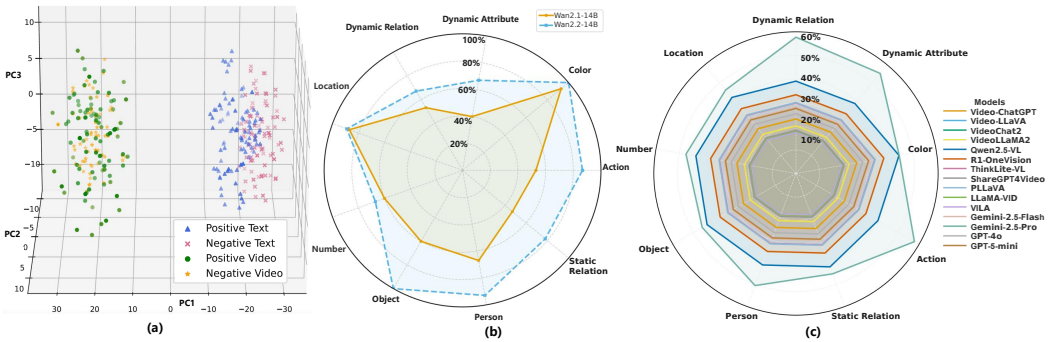

Figure 7: (***a***) t-SNE of Qwen2.5-VL-Instruct shows overlap between positive and negative samples for video and text pairs, indicating the model barely separates relation polarity across modalities. (***b***) Human verification rates for data synthesized by Wan 2.1 *vs.* Wan 2.2, where Wan 2.2 achieves higher success. (***c***) wACC on VIDPAIR-HALLUC for 15 models, illustrating wide performance variability, and Gemini-2.5-Pro achieves the best place.

Figure 8: Last-layer attention of (***a***) Qwen2.5-VL-Instruct and (***b***) ThinkLite-VL on adversarial text/video pairs with matched backgrounds and altered foreground actions. Heatmaps reveal focus patterns and alignment with correct *vs.* incorrect predictions.

Case Study. Following (Liu et al., 2025), we compare the mainstream open-source Qwen2.5-VL-Instruct with ThinkLite-VL—a video-reasoning model further fine-tuned from Qwen2.5—by analyzing last-layer attention. Contrary to the claim in (Liu et al., 2025), when the task reduces to binary judgments (Yes/No) without chain-of-thought requirements, the reasoning model consistently outperforms Qwen2.5-VL-Instruct. As shown in Figure 8 (b), ThinkLite-VL exhibits richer, better-localized last-layer attention and more reliably separates subtle differences between adversarial video pairs, leading to superior decisions. However, both models fail on adversarial text pairs, indicating that misleading high-level semantic perturbations can still derail their reasoning.

4 CONCLUSION

In conclusion, we introduce VIDPAIR-HALLUC, a novel benchmark for evaluating video hallucination in large video models under rigorously controlled conditions. By leveraging the PAIRFLOW pipeline, our dataset enables precise attribution of model errors to genuine hallucination by constructing adversarial video pairs with highly similar backgrounds but distinct foreground semantics. Extensive experiments reveal that both open-source and closed-source LVMs still struggle with robust and fine-grained video understanding, especially in adversarial scenarios. Our benchmark provides a new foundation for diagnosing and advancing video reasoning capabilities, highlighting the urgent need for more effective hallucination mitigation strategies. Beyond benchmarking, PAIRFLOW also provides a scalable data generation scheme. Its controllable, high-yield synthesis can be scaled up to supply diverse, hard negatives for training, thereby strengthening LVMs' video understanding and improving robustness in real-world deployments. We believe VIDPAIR-HALLUC will facilitate future research towards more reliable and trustworthy LVMs.

REFERENCES

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning large multimodal models for videos using reinforcement learning from ai feedback. *arXiv preprint arXiv:2402.03746*, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. *arXiv preprint arXiv:2503.14378*, 2025b.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2301–2310, 2024a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024b.

Sherry X Chen, Misha Sra, and Pradeep Sen. Instruct-clip: Improving instruction-guided image editing with automated data refinement using contrastive learning. *ArXiv*, abs/2503.18406, 2025.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418, 2024c.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. Vidhal: Benchmarking temporal hallucinations in vision llms. *arXiv preprint arXiv:2411.16771 , year=2024*.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24625–24634, 2024.

Xinpeng Ding, Kui Zhang, Jinahua 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.

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. *arXiv preprint arXiv:2405.21075*, 2024.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025a.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *arXiv preprint arXiv:2501.16629*, 2025b.

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.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023.

Google DeepMind. Nano banana – gemini image editing. https://gemini.google/overview/image-generation/, 2024. Accessed: 2025-09-22.

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

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xiaohu Jiang, Yixiao Ge, Yuying Ge, Dachuan Shi, Chun Yuan, and Ying Shan. Supervised fine-tuning in turn improves visual foundation models. *arXiv preprint arXiv:2401.10222*, 2024.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–13710, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. *arXiv preprint arXiv:2412.03735, year=2024*.

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

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*, pp. 22195–22206, 2024a.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. 2024b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2025.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.

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. *arXiv preprint arXiv:2311.10122*, 2023a.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023b.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*, 2025.

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*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. *arXiv preprint arXiv:2409.14750*, 2024c.

Yexin Liu, Zhengyang Liang, Yueze Wang, Muyang He, Jian Li, and Bo Zhao. Seeing clearly, answering incorrectly: A multimodal robustness benchmark for evaluating mllms on leading questions. *arXiv preprint arXiv:2406.10638*, 2024d.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*, 2024e.

Yue Ma, Xiaodong Cun, Sen Liang, Jinbo Xing, Yingqing He, Chenyang Qi, Siran Chen, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9385–9395. IEEE, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

OpenAI. Chatgpt 5. https://www.openai.com/, 2025. Accessed: 2025-09-24.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Sand-AI. Magi-1: Autoregressive video generation at scale, 2025. URL https://static.magi.world/static/files/MAGI_1.pdf.

Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. *ArXiv*, abs/2411.06686, 2024.

Achint Soni, Meet Soni, and Sirisha Rambhatla. Locatedit: Graph laplacian optimized cross attention for localized text-guided image editing. *ArXiv*, abs/2503.21541, 2025.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, Mengping Yang, Cheng Zhang, and Hao Li. Evalalign: Supervised fine-tuning multimodal llms with human-aligned data for evaluating text-to-image models. *arXiv preprint arXiv:2406.16562*, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2025. URL https://github.com/HumanSignal/label-studio. Open source software available from https://github.com/HumanSignal/label-studio.

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

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*, 2024a.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4581–4591, 2019.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025b.

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*, 2024b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. *ArXiv*, abs/2502.13928, 2025b.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. *arXiv preprint arXiv:2502.13928*, 2025c.

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. *arXiv preprint arXiv:2504.13169*, 2025d.

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

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

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*, pp. 9777–9786, 2021.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*, 2024.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 57240–57261. Curran Associates, Inc., 2024.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. *arXiv preprint arXiv:2501.00599*, 2024.

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

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024a.

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. *arXiv preprint arXiv:2402.13254*, 2024b.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024c.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10477–10486, 2023.

Ting Zhou, Daoyuan Chen, Qirui Jiao, Bolin Ding, Yaliang Li, and Ying Shen. Humanvbench: Exploring human-centric video understanding capabilities of mllms with synthetic benchmark data. *arXiv preprint arXiv:2412.17574*, 2024a.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024b.

Chenyang Zhu, Kai Li, Yue Ma, Longxiang Tang, Chengyu Fang, Chubin Chen, Qifeng Chen, and Xiu Li. Instantswap: Fast customized concept swapping across sharp shape differences. *arXiv preprint arXiv:2412.01197*, 2024.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2024.

A Related Work

A.1 Adversarial Video Pairs for Video Understanding

Building upon powerful LLMs and integrating various multimodal encoders, recent research has led to the development of MLLMs and LVMs (Liu et al., 2024b; Fu et al., 2025a; Yin et al., 2024; Wu et al., 2024; Li et al., 2023; Zhang et al., 2023; Lin et al., 2023a; Li et al., 2024a; Cheng et al., 2024; Jin et al., 2024; Li et al., 2025). Through supervised fine-tuning (SFT) on visual instruction-tuning data, these models have achieved impressive multimodal understanding and significantly improved human-computer interaction. However, inheriting the intrinsic hallucination issues of LLMs, LVMs frequently produce hallucinations or fail to align their understanding of visual content with human intuition (Liu et al., 2024a; Zhang et al., 2024c; Li et al.; Liu et al., 2024d). Increasing the scale of multimodal SFT data can alleviate these issues to some extent (Ahn et al., 2024; Tan et al., 2024; Jiang et al., 2024; Chen et al., 2024a; Yuan et al., 2024), but this approach is often constrained by high annotation costs and computational demands, especially in video scenarios where data and training requirements are substantially greater. To address these challenges, the community has introduced preference alignment techniques such as DPO (Rafailov et al., 2024), which align model outputs with human preference by leveraging pairs of preferred and rejected responses. Multimodal preference optimization extends this paradigm to visual and textual inputs, and has been widely adopted to enhance cross-modal alignment in MLLMs (Liu et al., 2024e; Xie et al., 2024; Zhou et al., 2024b; Wu et al., 2025d;b; Wang et al., 2024a). Recently, Hound-DPO (Zhang et al., 2024c) successfully applied multimodal DPO to LVMs, improving video understanding and mitigating hallucinations, yet still overlooked the alignment of visual inputs. (Huang et al., 2025; Ding et al., 2025) further employ the visual preference pairs to enhance video-language alignment. Notably, VistaDPO (Huang et al., 2025) introduces hierarchical preference optimization at the instance, temporal, and perceptive levels. The spatiotemporal adversarial video pairs within these preference pairs inherently contain subtle visual information, which enables VistaDPO to achieve significant improvement with relatively modest data volumes. However, the construction of such hierarchical preference pairs relies heavily on manual intervention to accurately annotate key frames and segments. This reliance, driven by the inherent complexity, fixed storylines, and spatiotemporal dependencies of real-world videos, significantly limits the scalability and efficiency of large-scale video preference data collection and makes annotation costly. Due to the high cost of obtaining real-world video preference pairs, most existing approaches synthesize large numbers of QA pairs from a single video and construct misleading, hallucinated questions to generate textual preference data for training (Zhang et al., 2024c; Huang et al., 2025) and evaluating models' multimodal hallucination capabilities (Wang et al., 2024b; Zhang et al., 2024a). Thus, the effectiveness of LVMs trained and evaluated on video pairs remains underexplored. Fortunately, advances in video generation technologies (Kong et al., 2024; Wang et al., 2025a; Sand-AI, 2025) combined with advanced editing techniques (Zhuang et al., 2024; Zhu et al., 2024; Soni et al., 2025; Chen et al., 2025; Shi et al., 2024) now enable precise control over spatial and temporal details, allowing the creation of fine-grained adversarial video pairs with plausible yet distinct variations. These adversarial pairs have the potential to enable fine-grained video-language alignment and robust evaluation of multimodal reasoning (Bai et al., 2025b). Building on this potential, this work focuses on leveraging adversarial video pairs specifically for evaluation. We propose PairFlow, a novel data pipeline designed for high-quality adversarial video pairs, as well as VidPair-Halluc for benchmarking LVMs for providing deeper insights into video hallucination.

A.2 Video Hallucination Benchmarks

LLMs gain strong language understanding from large-scale text pre-training, but their video encoders often lack similar representational strength. This mismatch can cause LLMs to generate confident yet incorrect outputs based on unreliable visual signals. To address this, researchers have introduced benchmarks that target various aspects of video hallucination for systematic diagnosis and mitigation. Early benchmarks (Yang et al., 2024; Zhang et al., 2024a) mainly evaluate model understanding within single videos. These works use spatio-temporal questions to test basic comprehension of object relations, actions, and event sequences in isolated visual settings. For example, Vript-HAL (Yang et al., 2024) introduces long, richly annotated videos and designs tasks that target action and object hallucinations. EventHallusion (Zhang et al., 2024a) emphasizes event-level reasoning to reveal biases rooted in language priors. Building on these foundations, later benchmarks introduce more adversarial textual contexts within single videos to test how robust models are against high-level

semantic disturbances. VideoHallucer (Wang et al., 2024b) follows this direction by systematically creating binary question pairs. Each pair includes *"one factual, one hallucinated"* question, covering both intrinsic and extrinsic hallucination types. This approach allows for detailed analysis of how models handle misleading or counterfactual queries. Recently, the evaluation paradigm has advanced further by using paired videos to better test a model's ability to align semantics and distinguish subtle visual differences. HallusionBench (Guan et al., 2024) and VIDHALLUC (Li et al.) both create video pairs that are visually similar but semantically different. HallusionBench focuses on disentangling language hallucination from visual illusion, using carefully selected image-question pairs to separate knowledge priors from visual evidence. VidHalluc explicitly targets temporal hallucinations by assembling over 5,000 video pairs, evaluating models on their ability to distinguish actions, temporal sequences, and scene transitions. It uses CLIP (Radford et al., 2021) and DINO (Caron et al., 2021) features to select semantically matched but visually diverse pairs, which directly exposes the limitations of current MLLMs in maintaining semantic consistency across related scenarios.

While these benchmarks have driven progress, the adversarial samples they use lack high visual similarity, leaving room for further improvement in sample quality. To further advance targeted, fine-grained hallucination evaluation, we enhance visual consistency across samples, enabling clearer assessment of a model's ability to capture subtle details in visual evidence.

B LIMITATIONS AND FUTURE WORK

Despite the promising results, our proposed PairFlow pipeline is currently constrained by the limited physical priors embedded in existing video generation models, which poses a bottleneck for further large-scale scaling. Additionally, this work primarily explores scenarios with a small number of video clips. Future research is needed to investigate the generation and evaluation of longer compositional videos involving more segments, which would better reflect real-world complexity. Overall, our benchmark holds significant potential for extension, providing valuable insights for more fine-grained video-language alignment and broader applications in advancing large model capabilities for video understanding.

C QUALITY ASSURANCE.

Quality Control. We employ trained participants to review all GPT-4.1-generated story scripts, ensuring each story is coherent and that all candidate endings are distinct yet contextually appropriate. Additionally, using Label Studio (Tkachenko et al., 2020-2025), annotators manually validate each generated video clip for consistency with its description and overall video quality, guaranteeing that all story videos meet our standards. Further details are provided in the supplementary material.

Human Validation of Adversarial Video Pairs. We recruit a team of annotators to manually review selected adversarial video pairs (*i.e.* Figure 4). Annotators are instructed to identify and eliminate pairs with the following issues: (***i***) insufficient or unclear foreground semantics in either video; (***ii***) adversarial pairs that do not exhibit highly similar backgrounds and significant differences in foreground semantics. This process ensures that only high-quality adversarial pairs are retained.

QAs Generation and Huamn Validation. For each validated video pair, we strictly generate one corresponding QA pair, covering 10 types of spatial and temporal hallucinations with targeted question templates. A separate group of annotators reviews all generated QAs, ensuring each question is relevant and accurately matches the video content. This process guarantees a challenging and reliable QA set for model evaluation.

D PROMPT FOR STORY GENERATION

D.1 ACTION SAMPLE GENERATION

Action Sample Generation

System Prompt
You are an expert dataset designer for visual action understanding. Your task is to generate action event samples in JSON format. Each sample must meet the following strict requirements:

1. **Action Focus:** Every sample describes a short, everyday scenario involving a human or animal performing a specific, observable action.
2. **Segments:** Each sample contains exactly three coherent segments (sentences), describing the action's beginning, middle, and end. The segments must be connected and form a natural mini-story.
3. **Candidate Values:** For each sample, provide exactly three candidate values (verbs or verb phrases) for the action, placed in the 'values' slot in each segment. - The three candidate values must correspond to **distinct, visually distinguishable external actions** (e.g., "open", "close", "wash"), not just differences in purpose or mental state. - Substituting any candidate value into the segments must result in a natural, logical, and visually clear story. - Avoid subtle or abstract distinctions (e.g., "compose"/"play" /"record"); prefer actions with clear physical differences.
4. **Formatting:** Output a JSON array called '"action"', where each element contains: - '"id"': a unique identifier (e.g., "action_0001")
- '"segments"': an array of three English sentences, each containing a 'values' placeholder
- '"values"': an array of three candidate action values (verbs or verb phrases)
5. **Diversity:** Ensure the actions, scenarios, and candidate values cover a wide range of everyday activities and are not repetitive.
6. **Language:** All content must be in fluent, idiomatic English.
7. **Sample Size:** Generate exactly 40 distinct samples per request.
Return only the JSON content as specified, with no additional commentary.

User Prompt
Please generate 40 action event samples according to the following requirements:
- Each sample should describe a short, everyday scenario in three connected English segments, with a 'values' placeholder for the action. - For each sample, provide three candidate action values (verbs or verb phrases) that are visually and physically distinct from each other, so that substituting any value results in a clear, observable difference in the described action. - Ensure the segments remain natural and logical with any candidate value substituted. - Output the results as a JSON array named '"action"', with each element containing '"id"', '"segments"', and '"values"' as described. - All content must be in English. - Generate exactly 40 samples.
Return only the JSON content.

D.2 Dynamic Visual Attribute Sample Generation

Dynamic Visual Attribute Sample Generation

System Prompt
You are a high-quality visual dynamic attribute sample generator. Your goal is to batch-generate samples for visual dynamic attribute scenarios according to the user's instructions. Each sample should include the following elements:
1. **ID**: A unique identifier (e.g., dynamic_attribute_0001).
2. **segments**: 3 descriptive sentences forming a coherent, intuitive visual story that depicts the dynamic change of an attribute, suitable for illustration via image/video.
3. **values**: 3 candidate values with significant distinction, perfectly fitting the story context. All values must represent clearly visible attribute changes (such as speed, quantity, distance, angle, color, shape, size, brightness, occlusion, density, etc.) and should avoid abstract or hard-to-visualize descriptions.
The generated samples must satisfy:
- Attribute change trends are clear and easy to visualize.
- Candidate values are highly distinguishable, with no ambiguous or hard-to-differentiate terms.
- Content is diverse and non-repetitive.
- Use concise, accurate English descriptions.
- Output must be in standard JSON structure, with the key named "dynamic_attribute" and the value being an array of samples.
Do not output anything except the JSON content.

User Prompt
Please batch-generate N high-quality samples (where N is specified by the user, e.g., 100) for a visual dynamic attribute recognition task. Each sample must include: - id: A unique identifier (e.g., dynamic_attribute_0001). - segments: 3 sentences forming a coherent, intuitive visual story describing the dynamic change of one attribute. - values: 3 candidate values that are highly distinguishable and clearly reflected in the scenario, tightly coupled with the story context.
The attribute types should cover speed, quantity, distance, angle, color, shape, size, brightness, occlusion, density, and other intuitive visual properties. Avoid abstract or hard-to-visualize descriptions.
The required output format is as follows:

```
{
  "dynamic_attribute": [
    {
      "id": "dynamic_attribute\0001",
      "segments": [
        "A runner starts on a city track, his speed {values}
        as he passes under streetlights.",
        "Midway through the race, his running pace {values},
        sweat visible on his brow.",
        "At the finish line, his speed {values}, and he raises
        his arms in triumph."
      ],
      "values": [
        "gradually increases",
        "gradually decreases",
        "remains steady"
      ]
    },
    ...
  ]
}
```

Please output only the JSON content, and do not include any text other than the JSON.

D.3 Sequence Sample Generation

Sequence Sample Generation

System Prompt
You are an expert data generator for visual storytelling datasets. Your task is to generate story sequence data for image or video generation. Each story consists of three segments describing a simple, everyday event or activity. For each story, provide three different orderings of the same three detailed actions, where each action includes vivid scene details suitable for visual generation. Return the result in JSON format as shown in the example.

User Prompt
Generate N story event samples as described above (replace N with the required number, e.g., 40). Each sample should have: - An ”id” field (e.g., ”sequence_0001”).
- A ”segments” array of three English sentences, each with a placeholder values[0], values[1], values[2] for the actions.
- A ”values” field: a 2D array with three permutations of the same three actions, each action being a richly detailed, visually descriptive phrase.
- All actions should fit naturally into the segment sentences and be suitable for generating realistic images or videos.

Example Format:

```
{
```

```
  "sequence": [
    {
      "id": "sequence_0001",
      "segments": [
        "The child first {values[0]}.",
        "Then, the child {values[1]}.",
        "Finally, the child {values[2]}."
      ],
      "values": [
        [
          "walks slowly across the grassy field, looking
          around curiously",
          "runs quickly past the playground, his arms swinging
          with excitepment",
          "jumps high over a small puddle, landing with a
          bright smile"
        ],
        [
          "jumps high over a small puddle, landing with a
          bright smile",
          "walks slowly across the grassy field, looking
          around curiously",
          "runs quickly past the playground, his arms swinging
          with excitepment"
        ],
        [
          "runs quickly past the playground, his arms swinging
          with excitepment",
          "jumps high over a small puddle, landing with a
          bright smile",
          "walks slowly across the grassy field, looking
          around curiously"
        ]
      ]
    }
    // ...more samples
  ]
}
```

Requirements: - Each story's three actions must be unique, richly detailed, and visually specific. - Each of the three "values" arrays must be a different permutation of the same three actions. - The language should be vivid and descriptive to support high-quality image/video generation. - Do not repeat actions or use generic verbs without scene context.
Generate N such samples in the specified format.

E Prompt for QA Generation

E.1 Multiple-Choice Question

Multiple-Choice Question Generation

System Prompt
You are an expert at generating multiple-choice question and answer (MCQ) data for multi-modal datasets. Given a video segment description, a question, and a set of answer choices, your task is to select the most appropriate answer based on the context. Always ensure your answer is accurate and based on the information provided in the segment description.

User Prompt
Given the following video segment description, question, and answer choices, select the single best answer. Respond only with the letter corresponding to the correct choice (e.g., "A").

Video Segment Description: {used_segment}
Question: {question}
Choices: A. {choice_A} B. {choice_B} C. {choice_C} D. {choice_D}

Instructions

1. Read the segment description carefully.
2. Choose the answer that best fits the context.
3. Only respond with the letter of the correct answer (e.g., **A**).

Example
Video Segment Description: The child picks up a balloon, ready to inflate it.
Question: What action was the child getting ready to do with the balloon?
Choices: A. inflate B. pop C. paint D. tie
Answer: A

E.2 Binary Question

Binary Question Generation

System Prompt
You are an expert at generating binary (Yes/No) question-answer data for multimodal datasets. Given a video segment description and a related Yes/No question, your task is to determine the correct answer based on the information provided. Your answer should be either "Yes" or "No", strictly according to the context.

User Prompt
Given the following video segment description and Yes/No question, answer with either "Yes" or "No" based on the context.

Video Segment Description: {used_segment}
Question: {question}

Instructions

1. Carefully read the segment description.
2. Answer the question strictly based on the information provided.
3. Respond only with "Yes" or "No".

Example
Video Segment Description: The child picks up a balloon, ready to inflate it.
Question: Did the child prepare to blow air into the balloon?
Answer: Yes

E.3 Open-Ended Question

Open-Ended Question Generation

System Prompt
You are an expert at generating open-ended question-answer (QA) pairs for multimodal datasets. Given a segment description from a video and a related question, your task is to generate a natural, contextually appropriate, and informative answer in English. The answer should accurately reflect the action or event described, and provide concise reasoning or context when appropriate. Use natural and fluent English, and avoid repetition.

User Prompt
Given the following video segment description and question, generate a natural, open-ended answer in English:

Video Segment Description: {used_segment}
Question: {question}

Instructions:

1. Your answer should be contextually relevant and reflect the action or event described in the segment.
2. Use natural, fluent English, and avoid repeating the question verbatim.
3. Provide a concise explanation or reasoning behind the action if possible.
4. Use the correct verb tense and pronouns based on the segment.
5. The answer should be a self-contained, informative sentence or short paragraph.

Example
Video Segment Description: The child picks up a balloon, ready to inflate it. She holds the balloon and inflates it with excitepment. Afterward, she smiles, having just inflated the balloon.
Question: What action was the child getting ready to do with the balloon?
Answer: The child was preparing to inflate the balloon, clearly excitepd about the activity. After inflating it, she smiled, showing her enjoyment of the moment.

F EDITING COMPARISON

Takeaway. As illustrated in Figure 10, the mainstream *commercial* editing models already deliver high, broad-spectrum performance, while prominent *open-source* options such as **Qwen-Image-Edit** reach respectable quality. These observations suggest that, as more advanced video generation/editing models mature, it will become practical to build high-quality *adversarial video pairs* on top of PAIRFLOW, with tight control over background and foreground semantics. Crucially, such pipelines are likely to be lower-cost and thus accessible to a wider range of labs and individual researchers.

G PERFORMANCE COMPARISON WITH RELATED BENCHMARKS

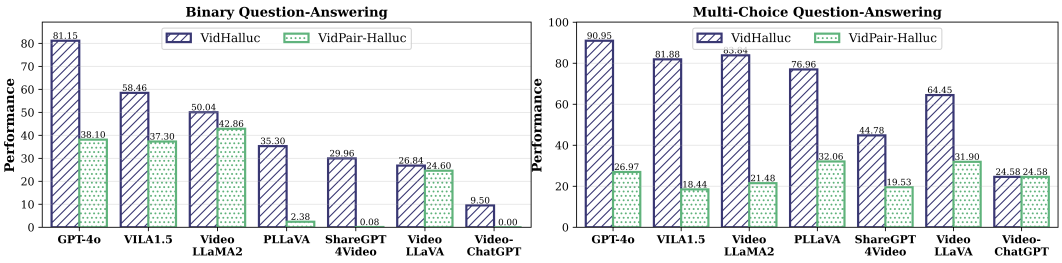

Figure 9: Performance comparison of VIDHALLUC *vs.* our VIDPAIR-HALLUC.

Takeaway. Unlike the previous benchmark VIDHALLUC, our proposed VIDPAIR-HALLUC further constructs adversarial video pairs with highly similar background semantics but significantly different foreground semantics. Based on this more challenging setting, we observe that most models suffer substantial performance degradation on VIDPAIR-HALLUC compared to VIDHALLUC as shown in Figure 9. This highlights the increased difficulty of accurately identifying fine-grained semantic differences when background cues are controlled, and underscores the necessity for more robust and nuanced video understanding capabilities in current LVMs.

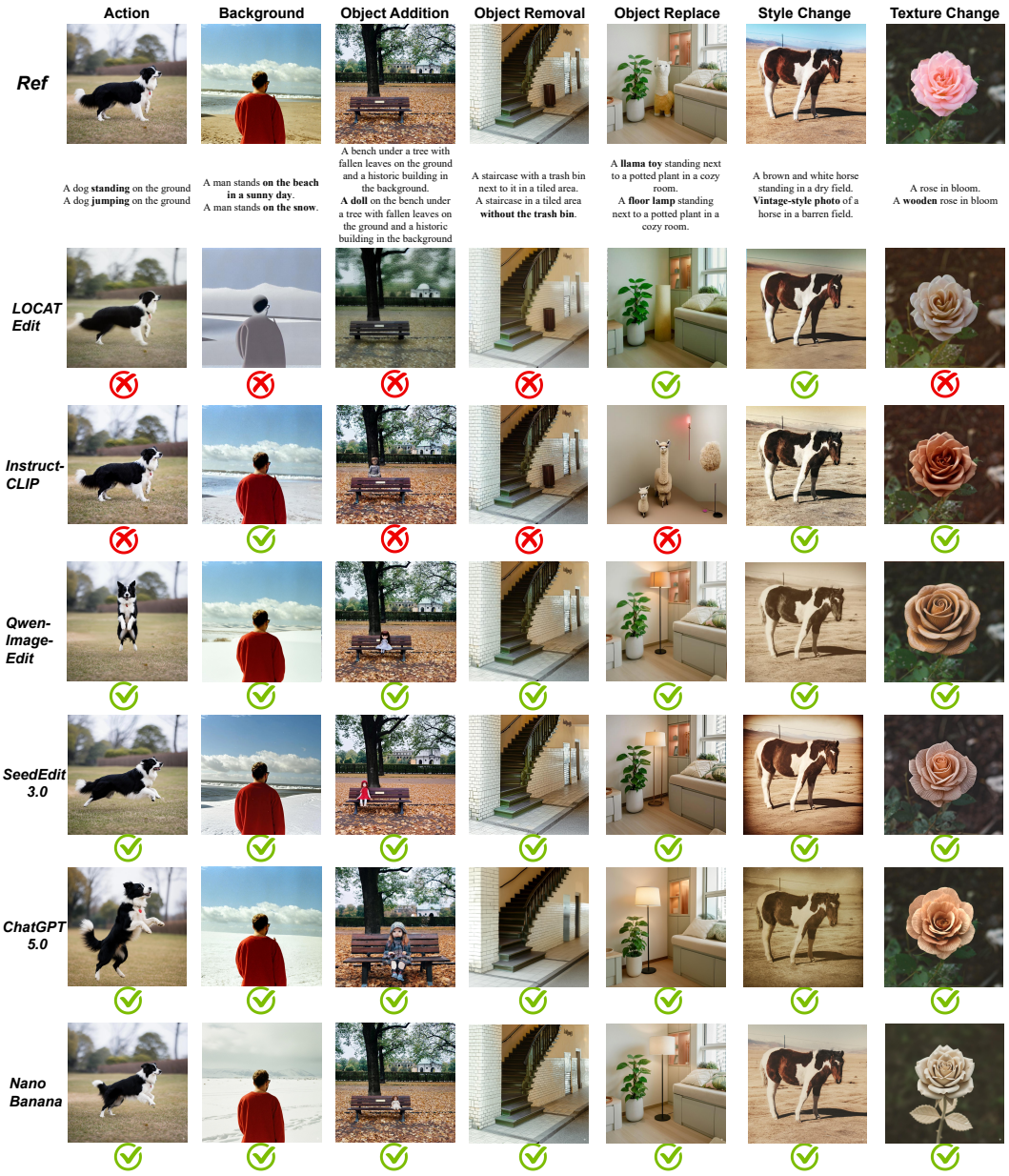

Figure 10: **Editing performance across methods.** The grid contrasts seven image-editing systems across seven edit types: *Action*, *Background*, *Object Addition*, *Object Removal*, *Object Replace*, *Style Change*, and *Texture Change*. The top row shows reference images and textual goals; subsequent rows report outputs from **LOCAT Edit** (Soni et al., 2025), **Instruct-CLIP** (Chen et al., 2025), **Qwen-Image-Edit** (Wu et al., 2025a), **SeedEdit 3.0** (Shi et al., 2024), **ChatGPT 5.0** (OpenAI, 2025), and **Nano Banana** (Google DeepMind, 2024), with green ticks and red crosses indicating success. Overall, **ChatGPT 5.0** and **Nano Banana** exhibit the most consistent successes, with **SeedEdit 3.0** also strong. **Qwen-Image-Edit** (open-source) achieves competitive quality, while **Instruct-CLIP** and **LOCAT Edit** lag on several edits (e.g., object manipulation and fine-grained style/texture control).