

No Place to Hide: Benchmarking Video Hallucination with Background-Controlled Pairs

Haojian Huang^{1,2} Harold Haodong Chen^{1,2} Meng Luo³
Junjia Du⁴ Shanqing Xu⁵ Ziheng Chen⁶
Yanxiang Huang⁷ Yinchuan Li² Ying-Cong Chen^{1,2*}

¹ The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

² Knowin AI, China ⁷ The Hong Kong Polytechnic University, Hong Kong SAR, China

³ National University of Singapore, Singapore ⁴ Nanyang Technological University, Singapore

⁵ Huazhong University of Science and Technology, Wuhan, China

⁶ University of Illinois Urbana-Champaign, Urbana, IL, United States

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. Evaluations on mainstream LVMs show persistent difficulty with robust fine-grained video understanding in adversarial settings, and code and data are available at the [project page](#).

1 Introduction

Progress in video understanding is driven by large-scale datasets [4, 13, 31, 51, 53, 59, 73] and advanced large video models (LVMs) [1, 5, 8, 30, 35, 41]. 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) [20, 64]. While LLMs undergo extensive pre-training, weaker video encoders lead to overconfident yet inaccurate outputs [10, 34]. To address video hallucination, benchmarks have evolved progressively. Initial works [15, 62, 65] 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

* Corresponding author.

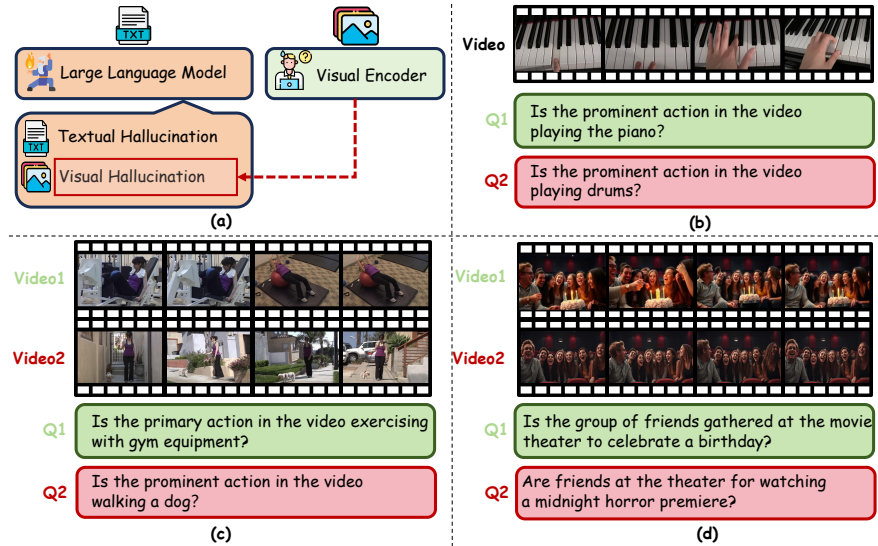


Fig. 1: (a) Hallucinations in LVMs, caused by the LLM or visual encoder, are categorized based on adversarial sources into textual or visual, where visual hallucination poses a greater challenge. (b) Adversarial question benchmarks (e.g., VideoHalluciner [53]) mainly induce textual hallucinations by perturbing the LLM. (c) VIDHAL-LUC [29] 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.

questions or distractor answers within single videos [53], 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 [18, 29] 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 se-

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 automated pipelines [17, 38, 56, 58, 66, 76], 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 contributions are as follows:

- We propose PAIRFLOW, a pipeline that constructs background-consistent, foreground-divergent adversarial video pairs for fine-grained video hallucination evaluation.
- 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 under rigorous background and foreground control.
- 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 Related Work

2.1 Video Hallucination Benchmarks

Video hallucination arises when an LVM produces confident responses that are not supported by the visual evidence. Early benchmarks mainly evaluate this behavior within a single video. Vript-HAL [62] and EventHallusion [65] ask spatio-temporal questions about object relations, actions, and event sequences, exposing failures in basic video grounding and event-level reasoning. Later work introduces stronger textual adversaries. VideoHalluciner [53], for example, builds paired binary questions in which one question is factual and the other is hallucinated, making it possible to diagnose whether a model follows misleading language priors rather than the video content.

Recent benchmarks further move from single-video evaluation to paired visual inputs. HallusionBench [18] disentangles visual evidence from language priors through carefully constructed image-question pairs, while VIDHALLUC [29] assembles thousands of video pairs to probe temporal hallucination. These paired settings are closer to the counterfactual nature of hallucination diagnosis, but they often rely on naturally different videos or feature-based retrieval. As a result, the visual background, camera motion, and foreground semantics may all vary at once. This makes it difficult to attribute a model error to the queried foreground fact alone. In contrast, VIDPAIR-HALLUC deliberately fixes the background as much as possible while changing the query-relevant foreground semantics, yielding a controlled setting that complements real-video benchmarks.

2.2 Adversarial Video Pairs for Alignment

Preference and contrastive alignment have become important tools for mitigating hallucination in MLLMs and LVMs. Textual preference optimization methods such as DPO [43] and multimodal variants [39, 50, 60, 74] align model outputs with preferred responses, and video-oriented methods such as Hound-DPO [67], PAMI-VDPO [11], and VistaDPO [20] demonstrate that visual preference pairs can improve video-language grounding. However, high-quality video preference pairs are expensive: real videos contain fixed storylines, complex temporal dependencies, and many non-target changes that require careful manual annotation.

Generative models provide a scalable alternative. Recent video generation and image/video editing systems [6, 27, 44–46, 49, 75, 76] make it increasingly practical to create fine-grained counterfactual samples with controlled local changes. Related work also spans multi-shot generation, multimodal affective understanding, cross-modal reliability, repository-dependency evaluation, temporal grounding, video repair, spatial reasoning, and embodied VLM robustness [12, 19, 21–25, 32, 55, 68–71]. This opens a path toward reusable hard negatives for curation, calibration, reranking, and contrastive or preference training. PAIRFLOW follows this direction, but uses synthesis primarily for evaluation: by controlling the non-target background and altering only the queried value or order, it isolates the hallucination source more cleanly than mined real-video pairs.

3 VidPair-Halluc: An Adversarial Benchmark for Video Hallucination

Existing benchmarks typically assess video hallucination either by posing misleading questions on a single video [9, 53, 65] or by pairing videos with low visual similarity but distinct semantics [7, 18, 29]. A more 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 [7, 18, 57]. However, constructing such pairs from real videos is costly and impractical, limiting current benchmarks’ ability to probe hallucinations under these nuanced

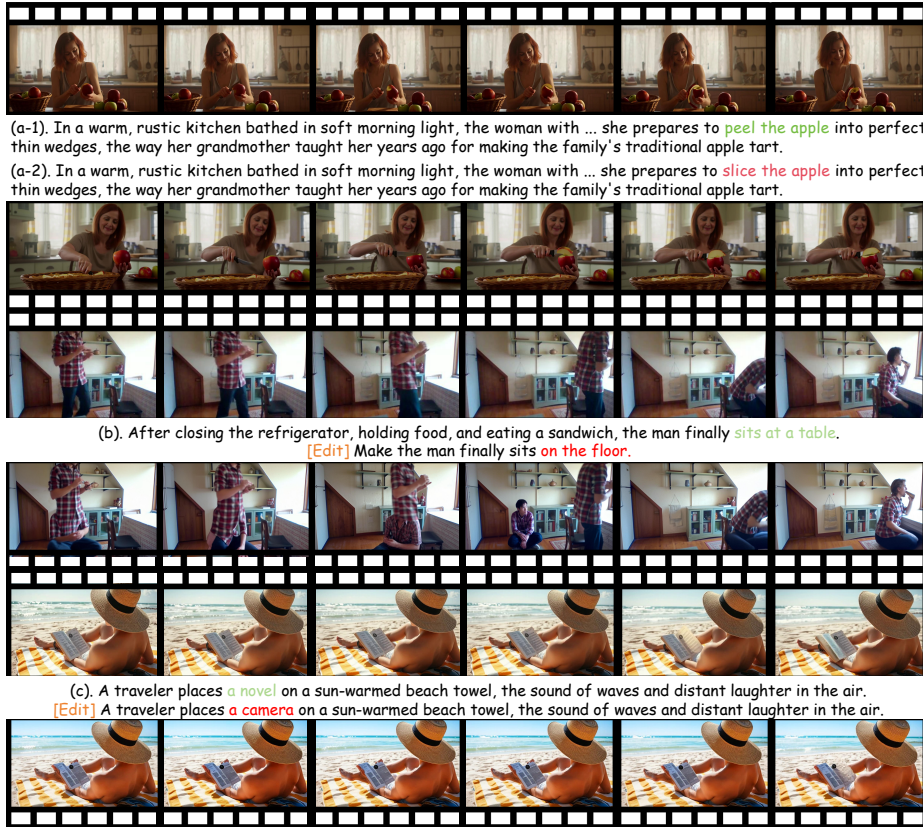


Fig. 2: Challenges in Generating High-Quality Adversarial Video Pairs. (a) Cutting-edge video generation models (*i.e.* HunyuanVideo [27]) create pairs with significant semantic differences but fails to maintain consistent visual backgrounds. (b) Advanced image editing methods (*i.e.* SeedEdit [45]), applied frame-by-frame, lead to noticeable inconsistencies. (c) Advanced video-editing techniques (*i.e.* TokenFlow [16]) result in pairs that lack faithfulness, stability and robustness.

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.

3.1 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 [66, 76], enabling a shift from hand-crafted corner cases to automated pipelines that efficiently scale up adversarial samples generation [38, 56, 58]. Motivated by this

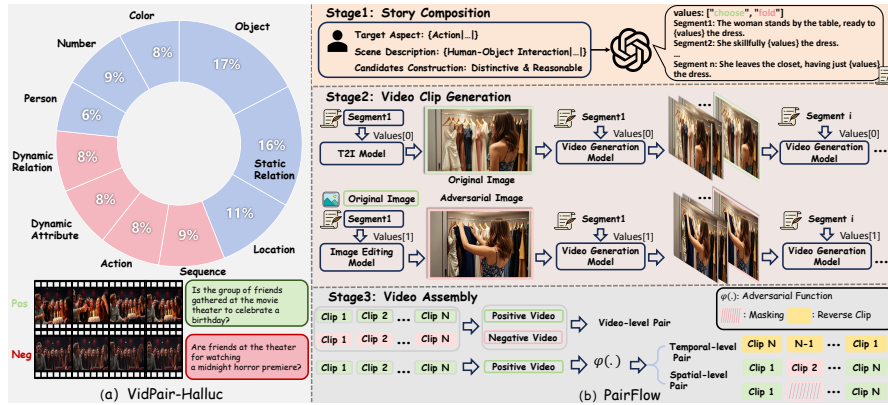


Fig. 3: (a) VIDPAIR-HALLUC evaluates hallucinations from temporal (33%) and spatial (67%) aspects. (b) PAIRFLOW generates video-text scenarios, synthesizes clips with controlled semantic variations, and assembles positive/negative pairs at multiple levels for robust hallucination benchmarking.

progress, we turn to recent advances in video generation and editing as a promising path to overcome the scalability bottleneck in the video domain.

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.* [3]) or editing frame by frame; (ii) applying video editing (*e.g.* [40, 72]); (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 [16], 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

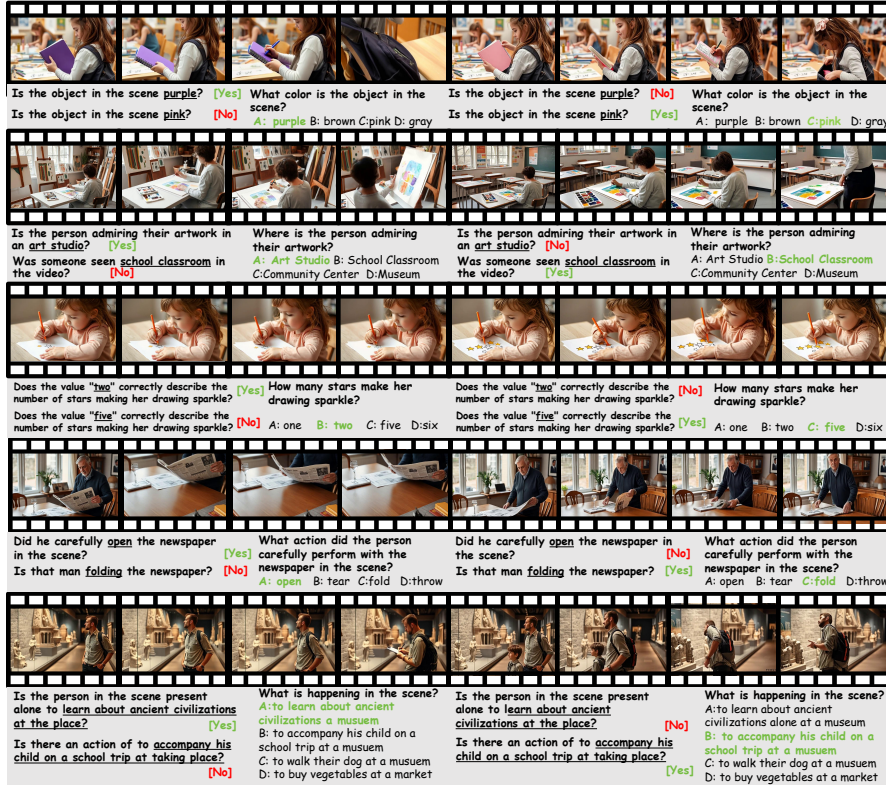


Fig. 4: Examples from VIDPAIR-HALLUC. Each row shows a positive-negative video-level pair, with four frames per video (left: positive, right: negative). The first three rows illustrate spatial reasoning, while the last two focus on temporal reasoning. Binary QA and MCQ are provided for each pair, highlighting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

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

Table 1: Comparison of existing video hallucination benchmarks.

Benchmark	Number of Question/Videos	Binary QA	Multi-Choice QA	Open-Ended QA	Visual Similarity	Control Pairs	Adversarial
HallusionBench [18]	1, 129/346	✓	✓	✗	✗	✓	✓
VideoHalluciner [53]	1, 800/948	✓	✗	✗	✗	✗	✓
Vript-HAL [62]	122/122	✗	✓	✗	✓	✗	✗
EventHallusion [65]	– /400	✓	✗	✓	✗	✗	✗
VidHALLUC [29]	9, 295/5, 002	✓	✓	✓	✗	✓	✓
VidPair-Halluc (Ours)	11, 523/2000	✓	✓	✓	✓	✓	✓

corresponding to segment V_i . This approach is mathematically denoted as:

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

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 .

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 $\varphi(\cdot)$ 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 $[V_N, V_{N-1}, \dots, V_1]$, 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.

Statistics. As shown in Table 1, the VIDPAIR-HALLUC dataset consists of 11,523 video-question pairs from 2,000 unique 15-second videos. While longer videos are available, we limit the length to 15 seconds to focus on manageable sequences and reduce challenges with longer content. Specifically, the dataset includes 4,000 binary QA, 4,000 multiple-choice questions (MCQ), and 3,523 open-ended QA, all of which have been verified for quality.

Our adversarial pairs are categorized into non-scene-dependent cases (e.g., object/action edits with the same background) and scene-dependent cases (e.g., person as background). In non-scene-dependent cases, we compute full-frame metrics, while in scene-dependent cases, we mask the consistent subject and compute metrics on the masked regions. As shown in Table 2, we use DINOv2, LPIPS, and SSIM metrics to assess background similarity. Evaluations are based on the keyframe at the start of each video clip, showing strong consistency across both types of cases and confirming high visual similarity across varying backgrounds.

Table 2: Background Similarity Metrics.

Metric	Random	Non-Scene	Scene (Masked)
DINOv2 ↑	0.21±0.08	0.94±0.02	0.81±0.05
LPIPS ↓	0.68±0.12	0.15±0.03	0.15±0.02
SSIM ↑	0.24±0.06	0.76±0.05	0.90±0.04

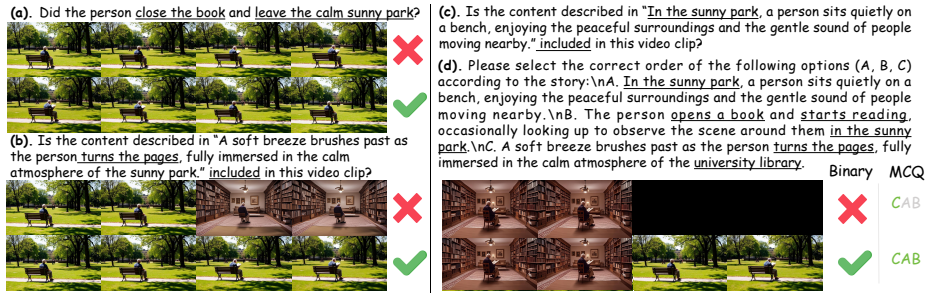


Fig. 5: Illustration of the increased challenge posed by temporal-level (a) and spatial-level (b) adversarial pairs compared to video-level adversarial pairs. Each spatial-level pair is accompanied by both binary QA (c) and MCQ (d) tasks. Notably, the MCQ format requires assessing and ranking the relevance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

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 [48] 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 samples with background drift, foreground ambiguity, identity or attribute drift, temporal incoherence, or visible artifacts. Finally, for every validated pair we instantiate targeted QA templates spanning ten spatial/temporal hallucination types, and a separate rater group confirms strict question-video alignment for all 11,523 QA instances. This four-level audit ensures that the generator never judges its own outputs and that retained errors can be traced to the intended foreground value or temporal-order change. Full protocols, rater instructions, and diagnostics are provided in the supplementary material.

3.2 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 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.

3.3 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).

3.4 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.

4 Experiment

4.1 Experimental Settings

Models. We evaluate 15 mainstream LVMs spanning open-source systems and closed-source APIs. The model set covers widely used video-language baselines, recent reasoning-oriented open models, and frontier proprietary models. Table 3 reports the full model list, citations, parameter sizes, and backbone details.

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 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

Table 3: Performance comparison of open-source and closed-source LVMs on 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** indicates best, underlined second best.

Method	Params	Language Model	Binary			Multi-Choice		Open-Ended
			wAcc ↑	FP (∼ 0)	Pct. Diff (∼ 0)	F1 ↑	vAcc ↑	Desc. ↑
<i>Open-source LVMs</i>								
Video-ChatGPT [41]	7B	LLaMA-7B	24.58	33.66	16.21	6.10	0.0	27.70
Video-LLaVA [35]	7B	Vicuna-7B-v1.5	31.90	35.30	23.26	53.52	24.60	34.74
VideoChat2 [31]	7B	Vicuna-7B-v0	18.47	<u>2.12</u>	-45.89	38.03	0.79	42.25
Video-LLaMA2 [8]	7B	LLaMA2-7B	21.48	43.00	33.41	62.91	42.86	40.85
Qwen2.5-VL-Instruct [2]	7B	-	<u>41.66</u>	48.71	14.77	61.50	42.86	45.07
R1-OneVision [63]	7B	-	35.52	27.24	<u>2.20</u>	29.23	28.33	42.37
ThinkLite-VL [52]	7B	-	31.86	33.40	4.79	43.29	37.27	48.39
ShareGPT4Video [5]	8B	LLaVA-Next-8B	19.53	1.67	-44.95	38.50	0.08	34.27
PLLaVA [61]	13B	Vicuna-13B-v1.5	32.06	45.72	-6.74	8.92	2.38	41.31
LLaMA-VID [33]	13B	Vicuna-13B-v1.5	18.79	53.13	49.75	33.80	15.08	30.52
VILA1.5 [36]	13B	-	18.44	15.18	-16.01	58.22	37.30	38.03
<i>Closed-source LVMs</i>								
Gemini-2.5-Flash [47]	-	-	29.83	18.97	-8.82	62.44	39.68	<u>50.23</u>
Gemini-2.5-Pro [47]	-	-	49.15	13.07	-2.12	67.32	<u>43.36</u>	54.68
GPT-4o [26]	-	-	26.97	29.16	10.97	59.15	38.10	47.89
GPT-5-mini [42]	-	-	29.33	19.65	3.28	<u>64.82</u>	45.28	49.33
Human	-	-	74.32	9.28	4.37	89.21	79.66	-

$\mathbb{V}_i$ 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:

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

where $b_{\mathcal{M}}(V, q)$ is a binary indicator of correctness for model $\mathcal{M}$ on video V and question q , and $\mathbf{1}(\cdot)$ 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 $\mathbb{V} = \bigcup_{i \in \mathcal{I}} \mathbb{V}_i$ be the set of videos across all adversarial pairs. We say a specific video $V \in \mathbb{V}_i$ 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:

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

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

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

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.

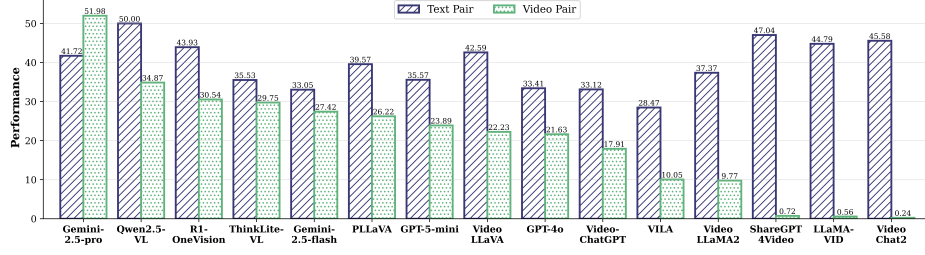


Fig. 6: Performance comparison between text-pair and video-pair hallucinations.

Formally,

$$d_y = \frac{\sum_{(V,q) \in \mathcal{A}} [\mathbf{1}(\mathcal{M}(V,q) = \text{yes}) - \mathbf{1}(y(V,q) = \text{yes})]}{|\mathcal{A}|}, \quad (5)$$

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.

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

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (6)$$

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

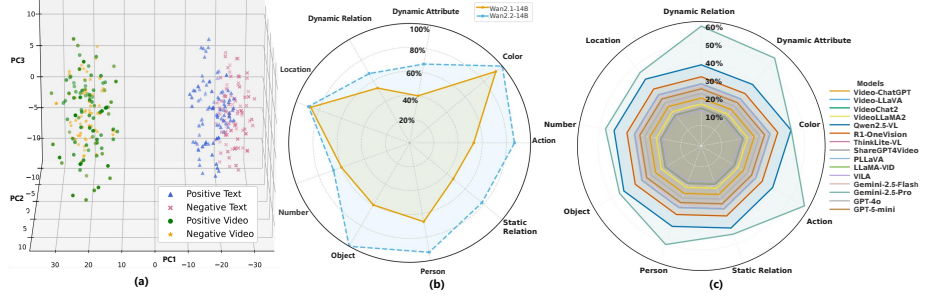


Fig. 7: (a) t-SNE of Qwen2.5-VL-Instruct shows overlap between positive and negative samples for video and text pairs, indicating the model barely separates relation polarity across modalities. (b) Human verification rates for data synthesized by Wan 2.1 *vs.* Wan 2.2, where Wan 2.2 achieves higher success.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

Implementation Details. For story script generation, we utilized the GPT-4.1 [26] API to automatically produce narrative descriptions. Based on these scripts, we employed the advanced open-source text-to-image model FLUX [28] to generate keyframes. Subsequently, precise foreground semantic editing was performed on

these keyframes using the advanced image editing model SeedEdit [45]. 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 [49]. All data processing and inference were conducted on a cluster of eight H100 GPUs. For open-ended questions evaluation, we employ GPT-4o [26] and follow the settings in [65]. 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.

4.2 Main Results

As illustrated in Figure 6, video-pairs with highly similar backgrounds present a more significant challenge for models compared to text-pairs, often exacerbating hallucinations. As shown in Table 3 and Figure 7 (c), closed-source LVMs outperform others, demonstrating superior calibration. Among these, Gemini-2.5-Pro stands out as the top performer, achieving the highest Binary wACC, the lowest FP, state-of-the-art Multi-choice results, and the most robust open-ended descriptions. GPT-5-mini, however, achieves the highest vAcc. The key advantage of low FP is its contribution to reducing spurious claims, an essential factor for safety-critical decision-making pipelines, where errors of commission can be costly. High wACC indicates broad resilience across adversarial video-text pairs and both clip- and video-level spatial semantics, rather than isolated successes. Among open-source models, Video-LLaMA2 excels in Multi-choice tasks (temporal reasoning), while Qwen2.5-VL achieves strong wACC, though with a higher FP, revealing a calibration gap. Overall, these results underscore the ongoing trade-off between spatial sensitivity and hallucination control. Bridging this gap will require synchronized advancements in temporal reasoning, spatial understanding, and risk-aware calibration.

4.3 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 from roughly 63% to about 83%, while Wan2.2 reaches approximately 70–100% pass@1 across the evaluated axes. Complementary editing results in Figure 10 suggest



Fig. 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.

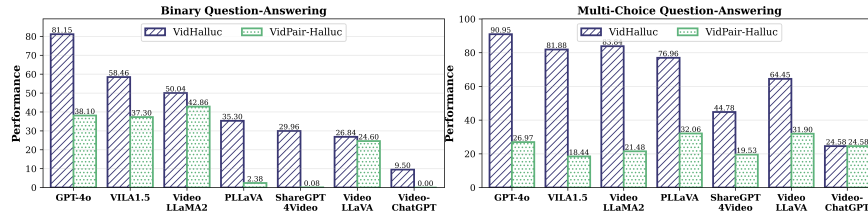


Fig. 9: Performance comparison of VIDHALLUC *vs.* our VIDPAIR-HALLUC.

that stronger image and video generators will make PAIRFLOW increasingly useful for both hallucination evaluation [3] and future contrastive or preference training [14, 20].

Case Study. Following [37], we compare Qwen2.5-VL-Instruct with ThinkLite-VL by analyzing last-layer attention. Contrary to the claim in [37], when the task reduces to binary judgments without chain-of-thought requirements, the reasoning model consistently performs better. As shown in Figure 8 (b), ThinkLite-VL exhibits richer, better-localized attention and more reliably separates subtle differences between adversarial video pairs, although both models remain vulnerable to adversarial text pairs.

Performance Comparison with Related Benchmarks. Unlike VIDHALLUC, VIDPAIR-HALLUC constructs adversarial video pairs with highly similar backgrounds but significantly different foreground semantics. As shown in Figure 9, most models suffer substantial degradation on VIDPAIR-HALLUC compared to VIDHALLUC [29], highlighting the difficulty of identifying fine-grained semantic differences when background cues are controlled.

Scope and limitations. Our benchmark complements real-video benchmarks: real videos capture natural diversity, while controlled pairs isolate queried foreground facts or temporal order for clearer attribution. The current release uses three 5-second clips per video; longer compositions remain future work.

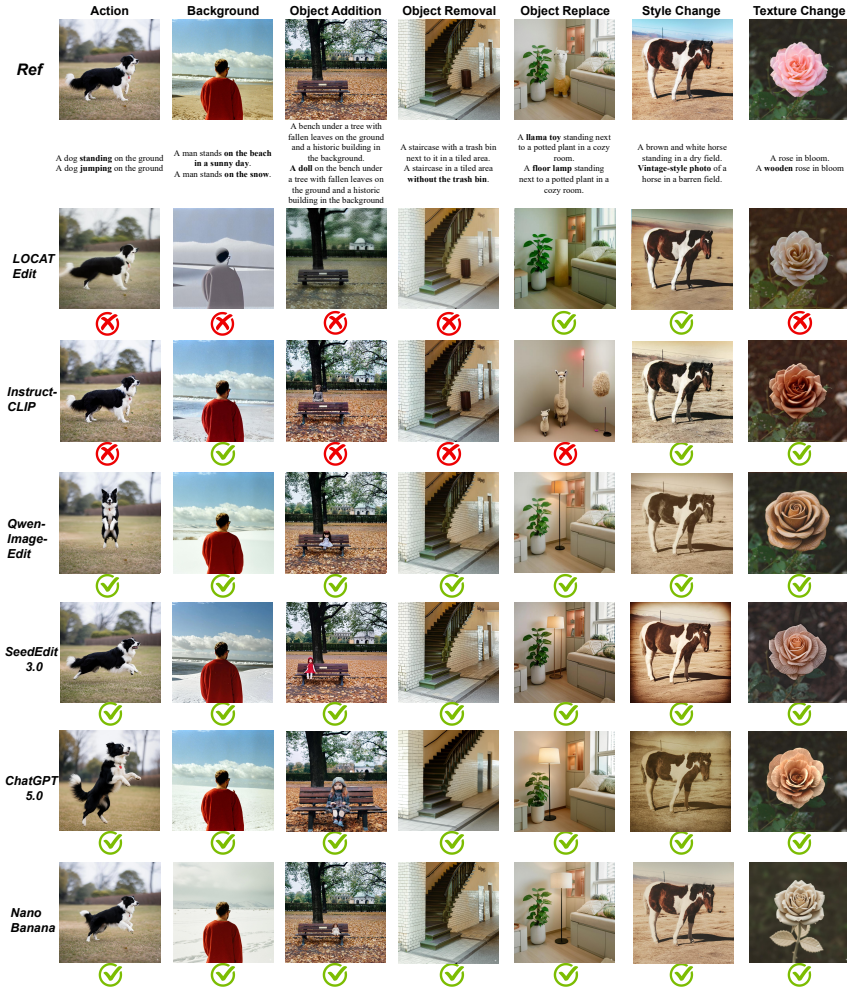


Fig. 10: Image editing performance comparison. We compare LOCAT Edit [46], Instruct-CLIP [6], Qwen-Image-Edit [54], SeedEdit 3.0 [45], ChatGPT 5.0 [42], and Nano Banana [17]; green/red marks indicate successful/failed edits. ChatGPT 5.0 and Nano Banana are the most consistent, followed by SeedEdit 3.0.

5 Conclusion

We introduce VIDPAIR-HALLUC, a background-controlled benchmark for diagnosing video hallucination with adversarial pairs. PAIRFLOW changes query-relevant foreground semantics while preserving backgrounds, enabling cleaner error attribution than mined videos. Results reveal persistent LVM weakness under fine-grained adversarial probes and motivate controlled pairs for diagnosis and alignment.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bai, Z., Ci, H., Shou, M.Z.: Impossible videos. arXiv preprint arXiv:2503.14378 (2025)
4. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)
5. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Tang, Z., Yuan, L., et al.: Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems* **37**, 19472–19495 (2024)
6. Chen, S.X., Sra, M., Sen, P.: Instruct-clip: Improving instruction-guided image editing with automated data refinement using contrastive learning. *ArXiv abs/2503.18406* (2025)
7. Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D., Chai, J.: Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems* **37**, 44393–44418 (2024)
8. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
9. Choong, W.Y., Guo, Y., Kankanhalli, M.: Vidhal: Benchmarking temporal hallucinations in vision llms. arXiv preprint arXiv:2411.16771 (2024)
10. Cui, X., Aparcedo, A., Jang, Y.K., Lim, S.N.: 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)
11. Ding, X., Zhang, K., Han, J., Hong, L., Xu, H., Li, X.: Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. arXiv preprint arXiv:2504.05810 (2025)
12. Du, J., Liu, Y., Guo, H., Wang, J., Huang, H., Ni, Y., Li, Z.: Dependeval: Benchmarking llms for repository dependency understanding. arXiv preprint arXiv:2503.06689 (2025)
13. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075 (2024)
14. Fu, J., Huangfu, S., Fei, H., Shen, X., Hooi, B., Qiu, X., Ng, S.K.: Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. arXiv preprint arXiv:2501.16629 (2025)
15. Gao, H., Qu, J., Tang, J., Bi, B., Liu, Y., Chen, H., Liang, L., Su, L., Huang, Q.: Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation. arXiv preprint arXiv:2503.19622 (2025)
16. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arxiv:2307.10373 (2023)
17. Google DeepMind: Nano banana: Gemini image editing. <https://gemini.google/overview/image-generation/> (2024), accessed: 2025-09-22

18. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacooob, Y., 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)
19. Hu, G., Xin, Y., Lyu, W., Huang, H., Sun, C., Zhu, Z., Gui, L., Cai, R., Cambria, E., Seifi, H.: Recent trends of multimodal affective computing: A survey from nlp perspective. arXiv preprint arXiv:2409.07388 (2024)
20. Huang, H., Chen, H., Wu, S., Luo, M., Fu, J., Du, X., Zhang, H., Fei, H.: Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. arXiv preprint arXiv:2504.13122 (2025)
21. Huang, H., Ma, K., Chen, J., Chen, H., Wu, Z., Zang, X., Fang, H., Ban, C., Sun, H., Chen, M., He, Z.: Adaptive evidential learning for temporal-semantic robustness in moment retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
22. Huang, H., Qiao, X., Chen, Z., Chen, H., Li, B., Sun, Z., Chen, M., Li, X.: Crest: Cross-modal resonance through evidential deep learning for enhanced zero-shot learning. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 5181–5190 (2024)
23. Huang, H., Qin, C., Li, Y., Chen, Y.C.: Find, fix, reason: Context repair for video reasoning. arXiv preprint arXiv:2604.16243 (2026)
24. Huang, H., Qin, C., Liu, Z., Ma, K., Chen, J., Fang, H., Ban, C., Sun, H., He, Z.: Trusted unified feature-neighborhood dynamics for multi-view classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 17413–17421 (2025)
25. Huang, H., Shi, J., Li, Y., Chen, Y.C.: Affordance agent harness: Verification-gated skill orchestration. arXiv preprint arXiv:2605.00663 (2026)
26. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
27. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
28. Labs, B.F.: Flux. <https://github.com/black-forest-labs/flux> (2024)
29. Li, C., Im, E.W., Fazli, P.: Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. arXiv preprint arXiv:2412.03735 (2024)
30. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
31. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., 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 (2024)
32. Li, Q., Yan, Q., Huang, H., Wu, P., Zhang, H., Zhang, Y.: Text-visual semantic constrained ai-generated image quality assessment. In: Proceedings of the 33rd ACM International Conference on Multimedia (2025)
33. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision (2024)

34. Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., Fu, Q., Liu, K.: A survey of multimodal large language models. In: Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering. pp. 405–409 (2024)
35. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
36. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoybi, M., Han, S.: Vila: On pre-training for visual language models (2023)
37. Liu, C., Xu, Z., Wei, Q., Wu, J., Zou, J., Wang, X.E., Zhou, Y., Liu, S.: More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. arXiv preprint arXiv:2505.21523 (2025)
38. Liu, J., Yang, X., Li, W., Wang, P.: Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. arXiv preprint arXiv:2409.14750 (2024)
39. Liu, Z., Zang, Y., Dong, X., Zhang, P., Cao, Y., Duan, H., He, C., Xiong, Y., Lin, D., Wang, J.: Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. arXiv preprint arXiv:2410.17637 (2024)
40. Ma, Y., Cun, X., Liang, S., Xing, J., He, Y., Qi, C., Chen, S., Chen, Q.: 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)
41. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
42. OpenAI: Chatgpt 5. <https://www.openai.com/> (2025), accessed: 2025-09-24
43. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36** (2024)
44. Sand-AI: Magi-1: Autoregressive video generation at scale (2025), https://static.magi.world/static/files/MAGI_1.pdf
45. Shi, Y., Wang, P., Huang, W.: Seedit: Align image re-generation to image editing. ArXiv **abs/2411.06686** (2024)
46. Soni, A., Soni, M., Rambhatla, S.: Locatedit: Graph laplacian optimized cross attention for localized text-guided image editing. ArXiv **abs/2503.21541** (2025)
47. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
48. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2025), <https://github.com/HumanSignal/label-studio>, open source software available from <https://github.com/HumanSignal/label-studio>
49. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

50. Wang, F., Zhou, W., Huang, J.Y., Xu, N., Zhang, S., Poon, H., Chen, M.: mdpo: Conditional preference optimization for multimodal large language models. arXiv preprint arXiv:2406.11839 (2024)
51. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: 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)
52. Wang, X., Yang, Z., Feng, C., Lu, H., Li, L., Lin, C.C., Lin, K., Huang, F., Wang, L.: Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. arXiv preprint arXiv:2504.07934 (2025)
53. Wang, Y., Wang, Y., Zhao, D., Xie, C., Zheng, Z.: Videohalluc: Evaluating intrinsic and extrinsic hallucinations in large video-language models. arXiv preprint arXiv:2406.16338 (2024)
54. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), <https://arxiv.org/abs/2508.02324>
55. Wu, L., Zhao, Y., Zhang, J., Chen, S., Chen, W., Chen, Z., Xu, T., He, Q., Hu, H., Huang, H., Wei, Y., Li, W., Li, Y., Chen, Y.C.: Robostressbench: Benchmarking vlm robustness to physical visual stress in embodied scenes. arXiv preprint arXiv:2606.00828 (2026)
56. Wu, S., Sun, F.Y., Wen, K., Haber, N.: Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. arXiv preprint arXiv:2502.13928 (2025)
57. Wu, S., Sun, F.Y., Wen, K., Haber, N.: Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. ArXiv **abs/2502.13928** (2025)
58. Wu, X., Guan, T., Li, D., Huang, S., Liu, X., Wang, X., Xian, R., Shrivastava, A., Huang, F., Boyd-Graber, J.L., et al.: Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. arXiv preprint arXiv:2406.10900 (2024)
59. Xiao, J., Shang, X., Yao, A., Chua, T.S.: 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)
60. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. arXiv preprint arXiv:2411.02712 (2024)
61. Xu, L., Zhao, Y., Zhou, D., Lin, Z., Ng, S.K., Feng, J.: Pllava: Parameter-free llava extension from images to videos for video dense captioning. arXiv preprint arXiv:2404.16994 (2024)
62. Yang, D., Huang, S., Lu, C., Han, X., Zhang, H., Gao, Y., Hu, Y., Zhao, H.: Vript: A video is worth thousands of words. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 57240–57261. Curran Associates, Inc. (2024)
63. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., Zhang, B., Chen, W.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)

64. Yuan, Y., Zhang, H., Li, W., Cheng, Z., Zhang, B., Li, L., Li, X., Zhao, D., Zhang, W., Zhuang, Y., et al.: Videorefer suite: Advancing spatial-temporal object understanding with video llm. arXiv preprint arXiv:2501.00599 (2024)
65. Zhang, J., Jiao, Y., Chen, S., Chen, J., Jiang, Y.G.: Eventhallusion: Diagnosing event hallucinations in video llms. arXiv preprint arXiv:2409.16597 (2024)
66. Zhang, J., Cai, M., Xie, T., Lee, Y.J.: Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. arXiv preprint arXiv:2402.13254 (2024)
67. Zhang, R., Gui, L., Sun, Z., Feng, Y., Xu, K., Zhang, Y., Fu, D., Li, C., Hauptmann, A., Bisk, Y., et al.: Direct preference optimization of video large multimodal models from language model reward. arXiv preprint arXiv:2404.01258 (2024)
68. Zhang, Y., Su, J., Huang, H., Chang, Y., Zhou, T., Xu, X., Xu, Y., Li, Y., Li, Z., Chen, Y.C.: Worldlines: Benchmarking and modeling long-horizon stateful embodied agents. arXiv preprint arXiv:2606.18847 (2026)
69. Zhao, P., Wang, L., Chen, C., Wang, P., Zhou, F., Li, R., Huang, H.: Star: Mitigating cascading errors in spatial reasoning via turn-point alignment and segment-level dpo. arXiv preprint arXiv:2604.00558 (2026)
70. Zhao, P., Wang, L., Wang, M., Chen, C., Zhou, F., Huang, H.: Evoempirbench: Dynamic spatial reasoning with agent-expver. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
71. Zheng, M., Xu, Y., Huang, H., Ma, X., Liu, Y., Shu, W., Pang, Y., Tang, F., Chen, Q., Yang, H., et al.: Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. arXiv preprint arXiv:2412.02259 (2024)
72. Zhou, S., Li, C., Chan, K.C., Loy, C.C.: Propainter: Improving propagation and transformer for video inpainting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10477–10486 (2023)
73. Zhou, T., Chen, D., Jiao, Q., Ding, B., Li, Y., Shen, Y.: Humanvbench: Exploring human-centric video understanding capabilities of mllms with synthetic benchmark data. arXiv preprint arXiv:2412.17574 (2024)
74. Zhou, Y., Cui, C., Rafailov, R., Finn, C., Yao, H.: Aligning modalities in vision large language models via preference fine-tuning. arXiv preprint arXiv:2402.11411 (2024)
75. Zhu, C., Li, K., Ma, Y., Tang, L., Fang, C., Chen, C., Chen, Q., Li, X.: Instantswap: Fast customized concept swapping across sharp shape differences. arXiv preprint arXiv:2412.01197 (2024)
76. Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: 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)