

CounterVid: Counterfactual Video Generation for Mitigating Action and Temporal Hallucinations in Video-Language Models

Tobia Poppi^{1,2,3}, Burak Uz Kent¹, Amanmeet Garg¹, Lucas Porto¹, Garin Kessler¹, Yezhou Yang¹, Marcella Cornia², Lorenzo Baraldi², Rita Cucchiara², Florian Schiffrers¹

¹Amazon Prime Video, ²University of Modena and Reggio Emilia, ³University of Pisa
{tobipop,burauzke,amanmega,lporto,kesslerg,imyzyang,floschi}@amazon.com,
{name.surname}@unimore.it, {name.surname}@phd.unipi.it

Abstract

Video-language models (VLMs) achieve strong multimodal understanding but remain prone to hallucinations, especially when reasoning about actions and temporal order. Existing mitigation strategies, such as textual filtering or random video perturbations, often fail to address the root cause: over-reliance on language priors rather than fine-grained visual dynamics. We propose a scalable framework for *counterfactual video generation* that synthesizes videos differing only in actions or temporal structure while preserving scene context. Our pipeline combines multimodal LLMs for action proposal and editing guidance with diffusion-based image and video models to generate semantic hard negatives at scale. Using this framework, we build CounterVid, a synthetic dataset of $\sim 26k$ preference pairs constructed from short counterfactual action clips and targeting both action recognition and controlled action-sequence ordering. We further introduce MixDPO, a unified Direct Preference Optimization approach that jointly leverages textual and visual preferences. Across Qwen2.5-VL and InternVL3 backbones, MixDPO substantially improves action recognition and temporal ordering, yields gains on most standard video hallucination benchmarks, and largely preserves general video understanding. Our source code, trained models, and dataset are available at <https://aimagelab.github.io/CounterVid>.

1 Introduction

Large-scale video-language models (VLMs) (Liu et al., 2023, 2024b; Chen et al., 2024; Zhu et al., 2023) have demonstrated strong performance on open-ended video understanding tasks (Maaz et al., 2024; Ding et al., 2024; Li et al., 2023; Lin et al., 2024a; Liu et al., 2024c; Zhang et al., 2024b). Despite this progress, these models remain susceptible to hallucinations, producing responses that are not supported by visual evidence (Liu et al., 2024a; Wang et al., 2024b; Bai et al., 2024).

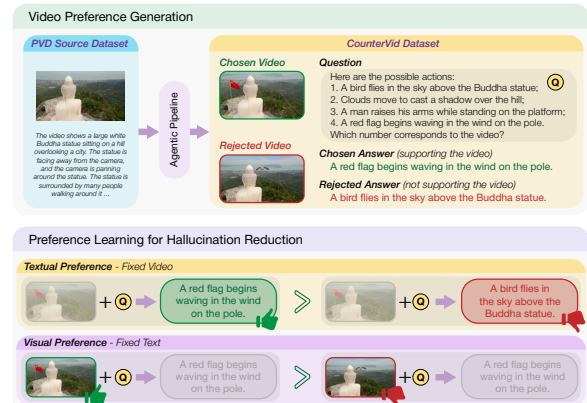


Figure 1: **Video preference generation and learning.** From a source video-caption pair, our pipeline creates semantically matched chosen/rejected videos and answers for textual and visual preference learning.

Video hallucinations manifest in diverse forms, ranging from static appearance errors, such as describing nonexistent objects or misidentifying attributes, to dynamic failures caused by incorrect temporal understanding. Recent studies (Wang et al., 2024b; Li et al., 2025a) show that modern VLMs often confuse visually similar actions or infer event sequences from language priors rather than observed motion. For example, a model may correctly recognize objects and scenes yet incorrectly assert that an action occurred, or assume a typical order of events even when contradicted by the video. Notably, such errors persist even in strong instruction-tuned models, highlighting limitations beyond surface-level reasoning.

Motivated by these observations, we focus on two critical and under-explored failure modes: *semantic action misidentification* and *incorrect inference of event order*. Existing approaches address video hallucination by introducing additional supervision, either through architectural modifications (Ma et al., 2024; Zhang et al., 2024a) or preference-based learning (Rafailov et al., 2023; Ding et al., 2025). However, many existing meth-

ods rely on negative samples from random perturbations, such as frame shuffling. While such perturbations disrupt temporal coherence, they rarely produce *semantic* counterfactuals, *i.e.*, videos that preserve scene context while differing meaningfully in actions or event order. Consequently, models may succeed by exploiting low-level artifacts or static cues, rather than learning fine-grained action dynamics (Huang et al., 2018).

To overcome these limitations, we introduce a scalable framework for *counterfactual video generation* that produces short action clips targeting action recognition and controlled action-sequence ordering. Instead of random augmentations, our approach constructs videos that differ only in their underlying action trajectories while preserving scene identity. The proposed pipeline integrates complementary generative components: multimodal LLMs (Minaee et al., 2024; Caffagni et al., 2024) for plausible alternative actions and structured edits, image editing models for action-completion frames (Wu et al., 2025; Labs et al., 2025), and image-to-video diffusion models (Wan Team et al., 2025) for generating temporally coherent clips. This modular design disentangles semantic reasoning from visual synthesis, while enabling precise control over the generated counterfactuals.

Using this framework, we compile **CounterVid**, a large-scale dataset of approximately 26k preference pairs built from short synthetic clips with controlled action variations. Fig. 1 summarizes the preference-data construction and learning setup. From each anchor scene, we generate multiple action-consistent clips and construct preference pairs targeting both action recognition and temporal ordering, spanning free-form, binary, and structured formats. These data support two complementary supervision signals: *textual preferences*, contrasting grounded and hallucinated answers under fixed visual input, and *visual preferences*, contrasting correct and counterfactual videos under fixed textual query and answer.

We leverage these signals through **MixDPO**, a simple framework based on Direct Preference Optimization (DPO) (Rafailov et al., 2023) that integrates textual and visual supervision. Its contribution is to formulate visual preferences over same-scene semantic counterfactuals and optimize them jointly with standard answer-side preferences. By jointly promoting output grounding and sensitivity to visual evidence, our approach directly targets the mechanisms underlying video halluci-

nation. Experiments on Qwen2.5-VL (Bai et al., 2025) and InternVL3 (Zhu et al., 2025) backbones show that fine-tuning with MixDPO yields substantial improvements in action recognition and temporal ordering, with positive transfer to established benchmarks such as EventHallusion (Zhang et al., 2024a) and VidHalluc (Li et al., 2025a).

Contributions. Our contributions are threefold:

- We introduce a modular counterfactual video generation pipeline that combines language-guided action proposal with diffusion-based generation to produce semantic hard negatives at scale.
- We construct CounterVid, a synthetic preference dataset designed to expose action-recognition and temporal-ordering hallucinations under controlled visual context.
- We present MixDPO, a unified preference-based alignment framework that jointly leverages textual and visual supervision to improve grounding and temporal sensitivity in VLMs.

2 Related Work

Video Hallucinations. Recent works have shown that VLMs frequently generate hallucinated outputs that are not grounded in visual evidence, despite strong performance on standard video understanding benchmarks (Bai et al., 2024; Liu et al., 2024a; Wang et al., 2024b). VideoHalluciner (Wang et al., 2024b) provides an initial taxonomy, distinguishing intrinsic hallucinations that misdescribe content present in the video from extrinsic ones that fabricate absent content. Subsequent studies emphasize temporal understanding as a particularly vulnerable dimension: models often confuse visually similar actions or infer event order from language priors rather than observed motion (Li et al., 2025a; Zhang et al., 2024a). Together, these works indicate that action recognition and temporal reasoning remain central challenges for VLMs.

Architectural and Decoding Strategies. Several approaches mitigate hallucinations by strengthening visual grounding through architectural or inference-time interventions. VISTA-LLAMA (Ma et al., 2024) enforces balanced attention over visual and textual tokens to reduce the tendency to ignore video input. Decoding-based methods (Zhang et al., 2024a; Leng et al., 2024), instead, counteract language priors by contrasting outputs generated from original and degraded or manipulated visual inputs. Self-refinement strategies, including Volcano (Lee et al., 2024), further encourage models to verify and

revise predictions against visual evidence (Wang et al., 2025; Guo et al., 2025). While effective, these methods typically require custom architectures, additional inference passes, or specialized decoding schemes, which can limit scalability.

Preference-Based Alignment. A different line of work reduces hallucinations through preference-based alignment. In the image domain, V-DPO (Xie et al., 2024) and mDPO (Wang et al., 2024a) extend DPO (Rafailov et al., 2023) with both output-level preferences (same input, different answers) and input-level preferences (different inputs, same query). In the video domain, PaMi-VDPO (Ding et al., 2025) constructs preference pairs using video augmentations such as frame shuffling, cropping, and temporal reversal, while related studies apply preference optimization on synthetic videos to reduce commonsense and physics-related hallucinations (Li et al., 2025b).

Despite their effectiveness, augmentation-based approaches are limited by the source content: random perturbations can disrupt temporal coherence but rarely produce counterfactuals that preserve scene context while varying actions or event order. As noted in (Ding et al., 2025), the most informative negatives are visually similar yet semantically distinct, which simple augmentations cannot reliably generate. Our work addresses this gap by synthesizing content-aware counterfactual videos with controlled action and temporal variations, and by jointly leveraging visual and textual preferences to promote visual grounding over language priors.

3 Proposed Method

We propose a scalable framework that (i) synthesizes *counterfactual* videos that differ only in action dynamics while preserving scene context, and (ii) aligns a VLM using a unified preference objective that couples *textual* and *visual* supervision.

3.1 Task Definition and Preference Data

Video QA with Hallucinations. Let V denote a video clip, Q a question, and A a free-form textual answer. A VLM with parameters θ defines a conditional distribution $\pi_\theta(A | V, Q)$. A *hallucination* occurs when A is not supported by the visual evidence in V , despite being linguistically plausible. Using short action clips and controlled concatenations of them, we focus on two dynamic failure modes: (i) **action recognition** (misidentifying the action in an otherwise correct response) and (ii)

temporal ordering (incorrectly inferring the order of concatenated actions). This controlled setting isolates order sensitivity.

Preference Supervision. Our training signal uses pairwise preferences of the form (x, y^+, y^-) or (x^+, x^-, y) , where x is the multimodal context, y is the response, and $+$ and $-$ superscripts indicate preferred and rejected alternatives. We construct two complementary preference types.

Textual Preferences (t-pref). For a fixed context $x = (V, Q)$, we prefer a grounded answer A^+ over a hallucinated answer A^- :

$$(x, y^+, y^-) = ((V, Q), A^+, A^-). \quad (1)$$

Visual Preferences (v-pref). For a fixed text pair (Q, A) , we prefer the *correct* visual context over a *counterfactual* one:

$$(x^+, x^-, y) = ((V^+, Q), (V^-, Q), A), \quad (2)$$

where V^+ and V^- depict different actions (or different action sequences) under nearly identical scene context. This directly penalizes “blind” generation, where $\pi_\theta(A | V^+, Q) \approx \pi_\theta(A | V^-, Q)$.

3.2 Counterfactual Video Generation Pipeline

We start from a large collection of real video-caption pairs $(V^{\text{real}}, C^{\text{real}})$ and use them only to extract an anchor frame; all preference videos are *synthetic*. Given one anchor frame, we synthesize a set of N action videos $\mathcal{G} = \{(V_i, C_i)\}_{i=1}^N$ that share the same static context (scene, objects, viewpoint) but differ in the action dynamics. These generated videos form the basis for both action-recognition and temporal-ordering preference pairs. An overview of the generation pipeline is shown in Fig. 2, and we elaborate on each part in the following sections.

① **Keyframe Extraction.** For each real pair $(V^{\text{real}}, C^{\text{real}})$, we select a representative keyframe I^{start} that best matches the caption and provides a stable starting state for editing and generation. We use a coarse-to-fine retrieval approach using a shared embedding space model (Zhai et al., 2023): (i) *coarse* sampling at 2 fps to identify a high-alignment temporal neighborhood, and (ii) *refined* sampling at 12 fps within that neighborhood to choose the final keyframe. We compute caption-frame similarity by averaging over multiple spatial crops (center/left/right) to reduce sensitivity to framing and to better capture the global scene.

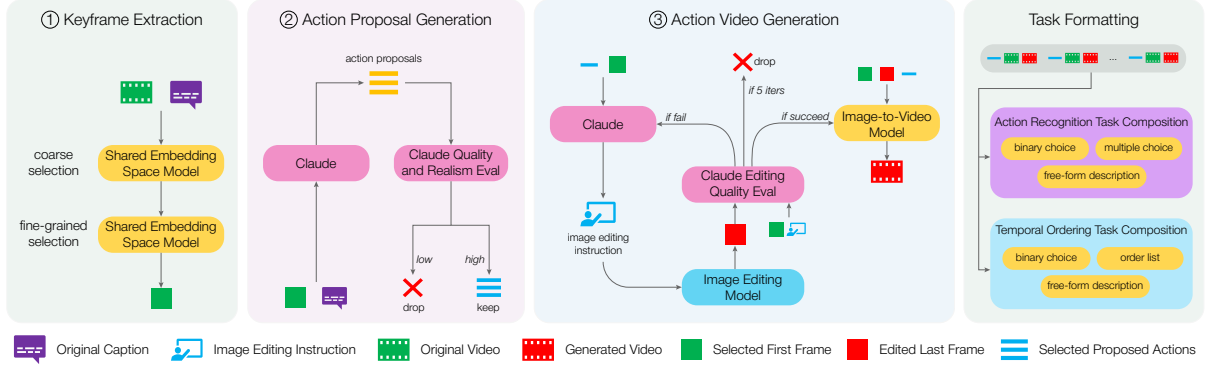


Figure 2: **Overview of the counterfactual generation framework.** Starting from a real video-caption pair, we extract a representative keyframe with a shared embedding space model, propose multiple alternative actions with a multimodal LLM, synthesize one counterfactual video per proposed action via image editing and image-to-video generation, and finally compose preference pairs for action recognition and temporal ordering.

② **Action Proposal and Filtering.** Conditioned on the keyframe I^{start} and the original caption C^{real} , we prompt a multimodal LLM (i.e., Claude (Anthropic, 2025)) to propose N alternative actions $\{C_i\}_{i=1}^N$ that are: (i) *plausible* given the scene state in I^{start} , (ii) *distinct* from each other (to ensure hard negatives), and (iii) *visually expressible* within a short clip. We then filter proposals using the same LLM with explicit criteria for realism, feasibility, and action clarity (templates are in Appendix D).

③ **Action Video Generation.** For each action caption C_i , we generate an *end frame* I_i^{end} depicting the completion of the action starting from I^{start} . We first convert C_i into structured editing instructions, then apply an image editing model (Wu et al., 2025) to obtain I_i^{end} . Because image editing can introduce semantic drift and artifacts, we employ an iterative refinement loop: after each attempt, the multimodal LLM evaluates whether I_i^{end} matches the intended action and preserves the scene identity; if not, it revises the editing prompt.

Finally, we synthesize a temporally coherent video V_i with an image-to-video model (Wan Team et al., 2025) conditioned on $(I^{\text{start}}, I_i^{\text{end}}, C_i)$. This yields a set of action-consistent clips $\{V_i\}_{i=1}^N$ that share the same context by construction (same anchor frame) and differ in the action trajectory. We further apply a post-generation safety screening step using visual safety classifiers. Samples flagged as potentially unsafe are removed from the dataset. More details are reported in Appendix A.

3.3 Task and Preference Construction

After generation, we discard $(V^{\text{real}}, C^{\text{real}})$, and construct samples for CounterVid exclusively from the generated set $\mathcal{G} = \{(V_i, C_i)\}_{i=1}^N$ obtained from the same anchor frame. Preference pairs are generated

for different target capabilities (i.e., action recognition and temporal ordering) and in different formats (free-form, binary, and multiple choice or order list depending on the sample type). In the following, let Q_{act} denote a generic action query (e.g., “What action is shown?”) and let $\mathcal{C} = \{C_i\}_{i=1}^N$ be the set of action captions proposed for the anchor.

3.3.1 Preference Pairs for Action Recognition

In action recognition, we employ a single generated clip V_i and ask the model to identify its action. To generate t-pref pairs, we keep the input context fixed, $x = (V_i, Q_{\text{act}})$, and contrast a grounded caption with a plausible but incorrect one sampled from the same anchor set:

$$A^+ = C_i, A^- = C_j, j \sim \text{Unif}(\{1..N\} \setminus i). \quad (3)$$

For v-pref pairs, we fix the text pair $(Q_{\text{act}}, A = C_i)$ and swap the video, so that the same answer is correct for one clip and false for the other:

$$(V^+, V^-) = (V_i, V_j), j \sim \text{Unif}(\{1..N\} \setminus i). \quad (4)$$

Formats. Free-form uses Q_{act} with A as caption. Binary choice presents one candidate and asks whether it matches the video. Multiple choice presents the candidate set $\mathcal{C}$, and the answer is the correct option index. In all cases, the preferred/rejected elements are defined via Eq. (3) or Eq. (4).

3.3.2 Preference Pairs for Temporal Ordering

Temporal ordering is evaluated under minimal contextual change. Both chosen and rejected sequences are formed by concatenating generated clips from the same anchor frame.

Sequence Construction. We sample K distinct actions $\{i_1, \dots, i_K\} \subset \{1..N\}$ and concatenate the

corresponding clips $S = [V_{i_1}; V_{i_2}; \dots; V_{i_K}]$, where $[\cdot; \cdot]$ denotes temporal concatenation. We define a *chosen* order S^+ as the sequence used to construct the concatenated video, and a *rejected* order S^- by applying a non-identity random permutation.

Preference Pairs Generation. For t-pref pairs, we keep the context fixed, $x = (S^+, Q_{\text{seq}})$, and contrast a grounded description with a hallucinated one obtained by caption swaps within the same anchor set. Concretely, A^+ describes the action order in S^+ , while A^- is formed by permuting or substituting one or more action captions from the same anchor set. For v-pref pairs, we fix the textual target to the correct order and swap the visual evidence between S^+ and S^- . This directly enforces that the model assigns a higher likelihood to the correct ordering when the underlying clips are permuted but visually similar in context.

Formats. Free-form uses Q_{seq} with A as an ordered list of actions. Binary choice presents a candidate ordering and asks whether it matches the video sequence. Order list presents the unordered action set and requires the model to output the correct temporal order as a sequence of indices.

Representative generated preference pairs from CounterVid are shown in Fig. 3.

3.4 The MixDPO Alignment Framework

We fine-tune a base VLM with a unified preference objective that combines textual and visual signals. Our goal is not only to improve answer quality, but to increase *input sensitivity*: answers should change when the visual evidence changes.

Preliminaries. DPO (Rafailov et al., 2023) optimizes a policy (*i.e.*, the video-language model being fine-tuned) π_θ given preference pairs (x, y^+, y^-) without an explicit reward model. Let π_{ref} be a frozen reference model, DPO defines an implicit reward as the log-likelihood ratio between the trainable policy and the reference model: $r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. The overall DPO objective is then defined as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[\log \sigma \left(\beta \left(r_\theta(x, y^+) - r_\theta(x, y^-) \right) \right) \right], \quad (5)$$

where σ is the sigmoid, and β controls the strength of the deviation from π_{ref} .

Unified MixDPO for Hallucination Mitigation. As multimodal hallucination often stems from the model ignoring the visual context, we define a unified loss that integrates two complementary signals

built on the two aforementioned preference types:

$$\mathcal{L}_{\text{MixDPO}}(\theta) = \mathcal{L}_{\text{t-pref}}(\theta) + \lambda \mathcal{L}_{\text{v-pref}}(\theta). \quad (6)$$

Here, we use $\lambda = 1$ as a balanced default between answer discrimination and visual grounding. Section 4.5 evaluates the sensitivity to this choice and to the textual/visual preference composition.

Textual Preference Loss. We force the model to sharpen its probability distribution over the correct tokens while suppressing plausible but visually unsupported sequences. Formally, given a textual preference triple $((V, Q), A^+, A^-)$, we define

$$\mathcal{L}_{\text{t-pref}} = -\mathbb{E} \left[\log \sigma \left(\beta \left(r_\theta((V, Q), A^+) - r_\theta((V, Q), A^-) \right) \right) \right]. \quad (7)$$

In CounterVid, A^- is constructed as an *action-caption swap* from the same anchor set, yielding a fluent but visually unsupported alternative, directly targeting language-prior guessing.

Visual Preference Loss. For a visual preference tuple $((V^+, Q), A, (V^-, Q), A)$, we define

$$\mathcal{L}_{\text{v-pref}} = -\mathbb{E} \left[\log \sigma \left(\beta \left(r_\theta((V^+, Q), A) - r_\theta((V^-, Q), A) \right) \right) \right]. \quad (8)$$

This term explicitly enforces that the same answer A should be assigned a high likelihood only under the correct visual evidence. When a model ignores video input, the two likelihoods become similar; maximizing their margin forces the policy to attend to discriminative temporal cues.

Minimizing $\mathcal{L}_{\text{MixDPO}}$ teaches the model both (i) *output grounding* (prefer grounded over hallucinated answers under fixed video) and (ii) *input sensitivity* (prefer the correct video over a counterfactual under fixed text), which is essential for mitigating temporal hallucinations in VLMs.

4 Experiments

4.1 Experimental Setup

CounterVid Details. The final version of CounterVid consists of 26,167 preference pairs spanning action recognition and temporal ordering, used to optimize VLMs with MixDPO. Action recognition includes 12,919 samples, distributed across free-form (4,416), binary-choice (4,416), and multiple-choice (4,087) formats. Temporal ordering consists of 13,248 samples, evenly split among free-form,

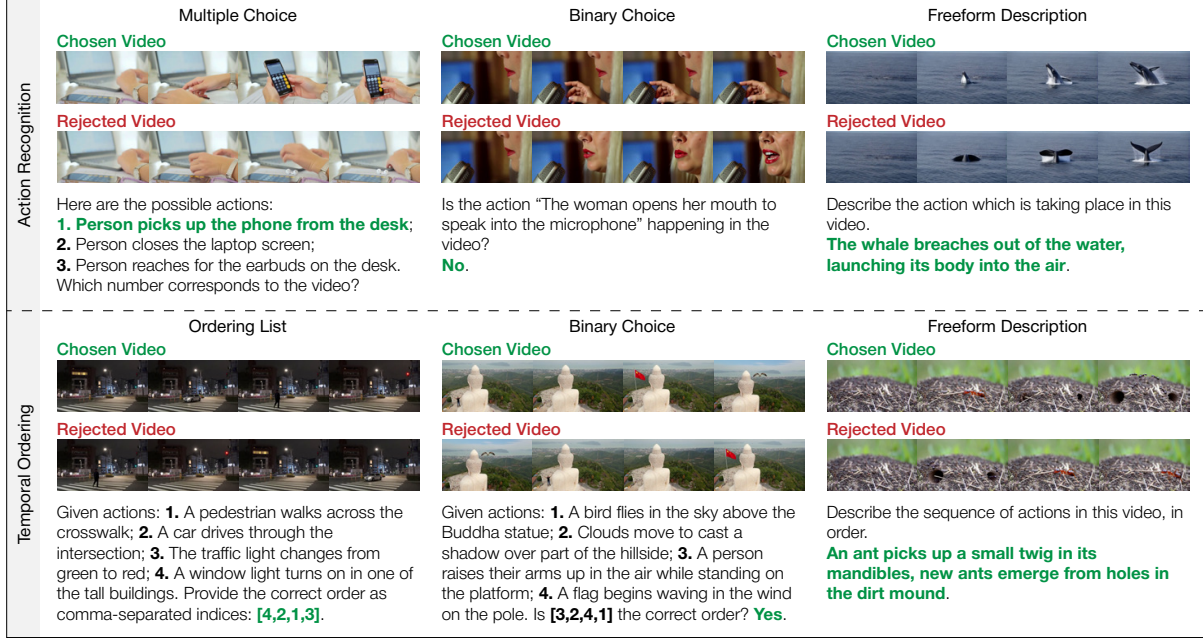


Figure 3: **Qualitative examples from CounterVid**, illustrating action recognition and temporal ordering samples across multiple-choice, order-list, binary, and free-form formats. All videos and answers are generated.

order-list, and binary-choice queries (4,416 each). We maintain a consistent preference ratio of 70% visual preferences (v-pref) and 30% textual preferences (t-pref). In addition to training data, we curate 2,910 held-out samples for benchmarking VLMs under the same evaluation settings. We perform the train-test split at the anchor level, ensuring preference pairs derived from the same anchor frame do not appear in both sets.

Data Generation Settings. We generate counterfactual videos using the pipeline described in Sec. 3.2. Real anchor frames and captions are selected from the PE Video Dataset (PVD) (Bolya et al., 2025), from which we only select clips shorter than 10 seconds, which are likely to contain a single action. Keyframe extraction is performed with SigLIP-SO400M (Zhai et al., 2023). Action proposal and filtering use Claude-4-Sonnet (Anthropic, 2025). We use Claude as one implementation of the multimodal planner/evaluator in our modular pipeline. Appendix B evaluates whether its filtering decisions are consistent with open-source VLM judges. End-frame generation is carried out via Qwen-Image-Edit (Wu et al., 2025) with an iterative refinement loop (up to five attempts) to ensure action correctness and scene consistency. Final videos are synthesized using Wan2.2-I2V-14B (Wan Team et al., 2025), producing clips of approximately three seconds at 680×384 resolution. Additional details are provided in the Appendix.

Baselines. We evaluate MixDPO across two VLM families: Qwen2.5-VL-3B and Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3-9B (Zhu et al., 2025). For each backbone, **Base** denotes the unfine-tuned model, **T-pref-DPO** applies DPO (Rafailov et al., 2023) using only textual preference pairs to isolate answer-level supervision, and **SFT** fine-tunes on original PVD captions (Bolya et al., 2025) to control for domain exposure. We also compare with **PaMi-VDPO** (Ding et al., 2025), a video-preference baseline whose chosen video and Crop/Shuffle negatives are jointly processed through its prompt-aware multi-instance objective. We use a matched three-epoch budget, and the main-table results use CounterVid videos as the chosen examples. To examine whether the observed behavior extends to a larger model, Appendix B.2 reports a matched CounterVid-only experiment on InternVL3-14B. Appendix A provides full implementation details, while Appendix B compares variants using PVD versus CounterVid chosen videos.

Evaluation Benchmarks. We evaluate our method on the held-out CounterVid split, a targeted in-family benchmark for action recognition and controlled short action-sequence ordering, and on a set of established video hallucination benchmarks. Specifically, we conduct experiments on VideoHalluciner (Wang et al., 2024b), VidHalluc (Li et al., 2025a), EventHallusion (Zhang et al., 2024a), and VideoHallu (Li et al., 2025b), which cover halluci-

Axis / Label	Free-Form	Binary	Multi-Choice	Order	All
<i>Label validity</i>					
Wrong	5.0%	6.9%	2.7%	7.0%	5.4%
Ambiguous	8.1%	4.9%	2.3%	11.9%	6.8%
Correct	86.9%	88.2%	95.0%	81.1%	87.8%
<i>Visual quality</i>					
Very Poor	2.0%	2.3%	2.4%	2.5%	2.3%
Poor	10.5%	9.5%	6.4%	12.4%	9.7%
Acceptable	21.4%	21.6%	21.7%	20.9%	21.4%
Good	58.1%	59.1%	62.2%	56.2%	58.9%
Excellent	8.0%	7.5%	7.3%	8.0%	7.7%
# Samples	108	103	46	43	300

Table 1: **Human evaluation by task format.** We assess label validity and visual quality separately to avoid conflating semantics with artifacts.

nations related to object existence, action recognition, temporal ordering, and event-level reasoning. To verify that hallucination mitigation does not degrade general video understanding, we additionally evaluate on VideoMME (Fu et al., 2025), NExTQA (Xiao et al., 2021), and TempCompass (Liu et al., 2024d). Further details on evaluation benchmarks and settings are provided in Appendix C.

4.2 Data Quality and Pipeline Evaluation

To assess the quality of the automatically generated CounterVid, we conduct a human evaluation on a representative subset of the held-out split. We evaluate two complementary axes: (i) preference-label validity (*i.e.*, whether the chosen video provides stronger visual support for the answer than the rejected one), and (ii) visual quality of the generated counterfactual videos.

Evaluation Protocol. We randomly sample 300 examples from the held-out set ($\approx 10.3\%$ of 2,910 samples), with proportional coverage across free-form (36.0%), binary-choice (34.3%), multiple-choice (15.3%), and order-list (14.3%) formats. Each sample is independently reviewed by ten evaluators, who inspect the generated videos together with the associated questions and answers. To evaluate video correctness, evaluators choose among correct, ambiguous, and wrong; for visual quality, they rate the video as excellent, good, acceptable, poor, or very poor. Additional annotation protocol details are given in Appendix A.

Results and Analysis. Table 1 reports the resulting breakdown. Overall, 87.8% of samples are judged correct and only 5.4% wrong, indicating that preference-label errors are uncommon. Under the quality axis, 66.6% of videos are rated good or excellent, and 88.0% are at least acceptable. Only

Model	Temporal Ord.			Action Rec.			
	FF	OL	BC	FF	MC	BC	Avg
▼ <i>Qwen2.5-VL-3B</i>	29.8	1.6	48.9	28.0	48.4	64.6	36.9
SFT	29.4	2.0	47.9	27.2	51.9	65.2	37.2
T-pref-DPO	32.2	15.7	56.0	29.6	58.7	69.5	43.6
PaMi-VDPO	32.9	1.4	47.5	29.1	52.3	68.0	38.5
MixDPO (Ours)	31.0	16.7	64.0	29.5	60.0	70.3	45.2
▼ <i>Qwen2.5-VL-7B</i>	70.3	16.5	57.8	55.1	68.6	78.2	57.8
SFT	69.3	14.9	55.9	55.3	67.7	79.2	57.1
T-pref-DPO	72.2	37.1	67.4	56.0	70.8	79.8	63.9
PaMi-VDPO	42.2	13.4	57.6	32.9	64.4	74.5	47.5
MixDPO (Ours)	72.2	43.8	72.7	56.3	71.2	80.7	66.2
▼ <i>InternVL3-9B</i>	66.3	13.2	53.0	49.9	65.3	75.4	53.8
SFT	66.2	16.9	51.7	53.2	69.7	74.8	55.4
T-pref-DPO	72.5	56.4	80.9	62.2	78.5	81.5	72.0
PaMi-VDPO	66.5	14.9	52.3	53.7	65.3	74.3	54.5
MixDPO (Ours)	74.1	61.3	84.7	63.2	79.6	81.1	74.0

Table 2: **CounterVid benchmark results.** Accuracy (%) on temporal ordering and action recognition tasks for Qwen2.5-VL and InternVL3. FF: free-form; OL: order list; BC: binary choice; MC: multiple choice.

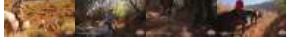
12.0% are poor or very poor, with severe degradation accounting for 2.3%. These results suggest that most generated counterfactuals are sufficiently interpretable for semantic preference learning, while the two-axis annotation makes label noise and visual artifacts explicit. Appendix B.3 analyzes the 46 unique samples flagged for an incorrect or ambiguous label, poor visual quality, or both, and Fig. 5 shows representative successes and failures.

4.3 Experimental Results

CounterVid Benchmark. Table 2 reports results on the held-out CounterVid benchmark across Qwen2.5-VL and InternVL3 models. MixDPO improves average accuracy from 36.9% to 45.2% on Qwen2.5-VL-3B, from 57.8% to 66.2% on Qwen2.5-VL-7B, and from 53.8% to 74.0% on InternVL3-9B. Gains are largest on temporal ordering, where predictions must track changed visual evidence: order-list accuracy rises from 1.6% to 16.7% for 3B and from 16.5% to 43.8% for 7B. On InternVL3-9B, MixDPO improves over T-pref-DPO by 2.0pp on average, with the largest gains on temporal ordering.

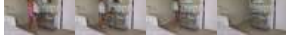
The SFT baseline yields minimal or negative changes (*e.g.*, -0.7pp for 7B), while PaMi-VDPO struggles on CounterVid, especially on temporal tasks, due to the lack of controlled action negatives. Beyond answer-level preference learning, MixDPO provides the largest additional gains on temporal ordering: on Qwen2.5-VL-7B, it improves over T-pref-DPO by +6.7pp on order-list and +5.3pp on binary temporal questions. This indicates that vi-

Question: Is horseback riding the predominant action captured in the video?



Qwen-2.5-VL-3B: **Yes**
T-pref only: **Yes**
Ours: **Yes**

A: picking up a backpack **B:** walking out the door **C:** applying lipstick **D:** closing a window



Qwen-2.5-VL-3B: **A**
T-pref only: **A**
Ours: **B**



Qwen-2.5-VL-3B: **Yes**
T-pref only: **Yes**
Ours: **No**

A: looking through the owner's manual **B:** jacking up the car **C:** cleaning the car's exterior **D:** checking the tire pressure



Qwen-2.5-VL-3B: **D**
T-pref only: **D**
Ours: **B**

Question: Is making a sandwich the predominant action captured in the video?



Qwen-2.5-VL-3B: **Yes**
T-pref only: **Yes**
Ours: **Yes**

Question: Below are two actions in the video: **A:** heating a pan. **B:** adding mashed potatoes and chutney, then mixing.



Qwen-2.5-VL-3B: **AB**
T-pref only: **AB**
Ours: **BA**



Qwen-2.5-VL-3B: **Yes**
T-pref only: **Yes**
Ours: **No**

Question: Below are two actions in the video: **A:** flying a drone **B:** jumping into the water.



Qwen-2.5-VL-3B: **AB**
T-pref only: **AB**
Ours: **BA**

Figure 4: **Qualitative hallucination correction.** On VidHalluc, MixDPO produces responses that better follow visual evidence than the base and text-only preference models, particularly for temporal-sequence questions.

Model	EventHal (Avg)	VidHal (TSH)	VideoHal (Overall)	VideoHcr (Avg)
▼ Qwen2.5-VL-3B	59.4	78.5	35.1	52.6
SFT	59.9	81.5	35.2	52.1
T-pref-DPO	64.3	80.0	36.3	53.6
PaMi-VDPO	66.4	76.2	36.3	50.7
MixDPO (Ours)	64.4	80.0	37.9	54.6
▼ Qwen2.5-VL-7B	70.1	82.7	38.2	54.1
SFT	66.7	82.7	39.0	53.9
T-pref-DPO	72.7	87.8	39.8	55.1
PaMi-VDPO	68.3	83.0	38.7	55.1
MixDPO (Ours)	72.5	90.5	39.3	55.7
▼ InternVL3-9B	64.6	39.3	26.5	54.6
SFT	64.0	46.5	27.4	55.6
T-pref-DPO	73.5	83.5	25.5	59.5
PaMi-VDPO	58.6	65.7	30.7	50.7
MixDPO (Ours)	73.7	88.8	25.7	58.9

Table 3: **Performance on established hallucination benchmarks.** Accuracy (%) across event-level, temporal, and content-based hallucination evaluations for Qwen2.5-VL and InternVL3 backbones.

sual preferences add input sensitivity when answer calibration alone is insufficient.

Established Hallucination Benchmarks. Table 3 reports results on four established hallucination benchmarks. MixDPO improves EventHallusion over the base model across all three backbones, reaching 64.4%, 72.5%, and 73.7% on Qwen2.5-VL-3B, Qwen2.5-VL-7B, and InternVL3-9B, respectively. The clearest additional gains over T-pref-DPO appear on VidHalluc TSH, with +2.7pp on Qwen2.5-VL-7B and +5.3pp on InternVL3-9B, while the 3B result is tied. Overall, across the 12 matched external comparisons, MixDPO improves over T-pref-DPO in eight cases and ties in one. The three regressions are small: -0.2 pp on EventHallusion and -0.5 pp on VideoHallu for Qwen2.5-VL-7B, and -0.6 pp on VideoHalluc for InternVL3-9B. Since these benchmarks differ from CounterVid in scene composition, action distribution, and failure modes, these results support positive but non-uniform transfer beyond our syn-

Model	VideoMME (Acc)	NExT-QA MC OE	TempCompass MC BC CM
▼ Qwen2.5-VL-3B	57.8	75.2 29.4	65.8 68.4 76.2
SFT	57.5	76.4 29.8	66.2 68.2 76.1
T-pref-DPO	57.4	76.8 29.6	66.8 69.0 77.5
MixDPO (Ours)	57.8	76.7 29.7	67.0 69.6 77.2
▼ Qwen2.5-VL-7B	61.4	74.7 32.2	73.2 75.5 76.2
SFT	61.5	75.5 32.2	73.4 75.6 79.6
T-pref-DPO	61.4	75.1 32.2	73.4 75.8 77.6
MixDPO (Ours)	61.5	74.9 32.2	73.7 75.9 80.2
▼ InternVL3-9B	64.8	80.7 33.2	69.3 70.9 76.7
SFT	64.5	80.8 33.0	69.8 72.4 76.4
T-pref-DPO	64.9	80.9 33.3	70.3 74.2 70.7
MixDPO (Ours)	64.7	80.9 33.0	70.1 74.4 69.5

Table 4: **General video benchmarks.** Accuracy (%) on VideoMME, NExT-QA, and TempCompass for Qwen2.5-VL and InternVL3. MC: multiple choice; OE: open-ended; BC: binary choice; CM: caption matching.

thetic benchmark, with the strongest gains on the temporal hallucination settings most directly targeted by our method.

General Video Understanding. Table 4 shows that general video understanding is largely preserved. VideoMME remains stable across Qwen2.5-VL and InternVL3, while NExT-QA improves for Qwen2.5-VL-3B and remains close to baseline for 7B and 9B. TempCompass improves for Qwen2.5-VL, including a caption-matching gain from 76.2% to 80.2% for 7B, but its caption-matching subset decreases for InternVL3. All benchmarks use their official splits without duration filtering; VideoMME therefore includes short, medium, and long videos. Although our method does not explicitly target long-form reasoning, evaluation on the general benchmarks tests its behavior beyond the short clips used for supervision.

4.4 Qualitative Analysis

Fig. 4 presents representative VidHalluc examples. The base and text-only preference models often favor plausible but unsupported actions or event

Model	Setting	TempOrd	ActRec	Avg.
Qwen2.5-VL-7B	$\lambda=1$, 70/30	62.9	69.3	66.1
	$\lambda=0.5$	62.9	69.0	66.0
	$\lambda=2.0$	62.3	69.1	65.7
	v/t = 50/50	62.7	69.3	66.0
	v/t = 90/10	64.2	68.8	66.5
InternVL3-9B	$\lambda=1$, 70/30	73.2	74.7	74.0
	$\lambda=0.5$	73.1	74.9	74.0
	$\lambda=2.0$	73.0	74.6	73.8
	v/t = 50/50	72.0	74.8	73.4
	v/t = 90/10	74.1	74.7	74.4

Table 5: **Loss-weight and preference-mixture sensitivity.** Controlled CounterVid configurations varying one factor around $\lambda=1$ and v/t = 70/30.

sequences, whereas MixDPO produces responses that better follow the video content and correct both action and temporal hallucinations. Beyond correcting explicit errors, we observe that MixDPO exhibits more cautious behavior when the visual evidence is ambiguous or partially occluded, avoiding confident but unsupported claims. This qualitative behavior complements the quantitative results, suggesting that mixed visual and textual preference learning improves both visual grounding and sensitivity to action order.

4.5 Robustness Analysis

Table 5 shows that MixDPO is not sensitive to a narrow balance between preference types. Changing λ fourfold alters average accuracy by at most 0.4pp on Qwen2.5-VL-7B and 0.2pp on InternVL3-9B. Moving from a 50/50 to 90/10 visual/textual mixture changes the average by 0.5pp and 1.0pp, respectively. Although 90/10 is highest overall, it exceeds 70/30 by only 0.4pp on each backbone and does not improve every task group. We treat the sweep as a robustness analysis rather than selecting a new mixture on the same evaluation benchmark, retaining 70/30 as the balanced configuration used throughout. Appendix B.1 reports all six submetrics and the complementary β sweep.

5 Conclusion

We presented a scalable framework for mitigating hallucinations in VLMs by coupling counterfactual video generation with unified preference-based alignment. By synthesizing semantic hard negatives that preserve scene context while varying action dynamics, our approach directly targets action misidentification and temporal-ordering errors that commonly arise from over-reliance on language priors. Across Qwen2.5-VL and InternVL3

backbones, CounterVid and MixDPO yield their strongest gains on these targeted capabilities, improve most external hallucination benchmark comparisons, and largely preserve general video understanding. This work highlights the promise of generative counterfactual data as a practical mechanism for improving visual grounding in VLMs, while leaving long-form temporal reasoning and broader component replacement for future work.

Limitations

While our approach achieves strong performance, several aspects offer opportunities for future extension and refinement. First, the evaluation mechanisms used for action proposal filtering and video generation could be further refined, for example, through more comprehensive sanity checks, to increase the proportion of high-quality samples in the final dataset. Second, the quality of the generated counterfactuals depends on the underlying generative models and is expected to improve as image and video generation methods continue to advance. Residual synthesis artifacts may also provide shortcut signals; our quality controls and external transfer results make a purely artifact-driven explanation less likely, but do not eliminate this confound. Third, our current counterfactuals focus on short actions (≈ 3 s) and controlled concatenations of those clips. Although we evaluate official external splits without duration filtering, including VideoMME’s short, medium, and long videos, the method does not directly address long-range causal dependencies, overlapping events, or multi-agent interactions. Fourth, the pipeline is modular at the interface level and alternative VLM judges reproduce most filtering decisions, but we do not establish invariant quality under replacement of the image editor or video generator; cross-generator validation remains future work. Finally, we adopt parameter-efficient fine-tuning by freezing the vision encoder; while full end-to-end training may further enhance sensitivity to fine-grained visual cues, it incurs substantially higher computational cost and optimization complexity, which we leave for future exploration.

Ethical Considerations

The ability to synthesize context-aware alternative actions via our counterfactual generation framework is a dual-use technology; while designed for model alignment, it could potentially be repurposed

to create misinformation or deceptive video content. We acknowledge that releasing our methodology and code inherently carries this risk, yet we prioritize the scientific necessity of addressing the widespread problem of VLM hallucinations. In addition, because our framework relies on large-scale generative models, the resulting synthetic data may inherit biases or stereotypes present in those models. To reduce harmful content in the released dataset, we screen generated videos with automated visual safety classifiers (NudeNet (Bedapudi, 2019) and Q16 (Schramowski et al., 2022)) and remove the approximately 0.6% of samples flagged as potentially unsafe. The dataset is derived from PVD only through anchor frames; it does not include source metadata and follows the privacy protections of the original source data. Finally, we note that large-scale synthetic video generation has environmental costs; by providing our code to the community, we enable other researchers to build upon our findings and alignment methodology without the need for redundant, large-scale computational trials.

Acknowledgments

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara acknowledge funding from the EU Horizon project ELLIOT (GA No. 101214398).

References

- Anthropic. 2025. [System Card: Claude Opus 4 & Claude Sonnet 4](#).
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, and 8 others. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2404.18930*.
- Praneeth Bedapudi. 2019. NudeNet: Neural Nets for Nudity Classification, Detection, and Selective Censoring. <https://github.com/notAI-tech/NudeNet>. GitHub repository.
- Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. 2025. Perception Encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*.
- Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. In *ACL Findings*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *CVPR*.
- Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. 2024. Holistic Autonomous Driving Understanding by Bird’s-Eye-View Injected Multi-Modal Large Models. In *CVPR*.
- Xinpeng Ding, Kui Zhang, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. 2025. PaMi-VDPO: Mitigating Video Hallucinations by Prompt-Aware Multi-Instance Video Preference Learning. *arXiv preprint arXiv:2504.05810*.
- Chaoyou Fu, Yuhao Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-Modal LLMs in Video Analysis. In *CVPR*.
- Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. 2025. TRACE: Temporal Grounding Video LLM via Causal Event Modeling. In *ICLR*.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.
- De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. 2018. What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets. In *CVPR*.
- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025.

- Flux.1 kontekst: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.
- Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision. In *NAACL*.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *CVPR*.
- Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025a. Vid-Halluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding. In *CVPR*.
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. VideoChat: Chat-centric Video Understanding. *arXiv preprint arXiv:2305.06355*.
- Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Fuxiao Liu, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. 2025b. VideoHallu: Evaluating and Mitigating Multi-modal Hallucinations on Synthetic Video Understanding. *arXiv preprint arXiv:2505.01481*.
- Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *EMNLP*.
- Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024b. Evaluating Text-to-Visual Generation with Image-to-Text Generation. In *ECCV*.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A Survey on Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2402.00253*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved Baselines with Visual Instruction Tuning. In *CVPR*.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024d. TempCompass: Do Video LLMs Really Understand Videos? *arXiv preprint arXiv:2403.00476*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.
- Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiaoshi Feng, and Yi Yang. 2024. VISTA-LLAMA: Reducing Hallucination in Video Language Models via Equal Distance to Visual Tokens. In *CVPR*.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *ACL*.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. *arXiv preprint arXiv:2402.06196*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*.
- Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. 2022. Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content? In *ACM FAccT*.
- Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, and 43 others. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.
- Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mDPO: Conditional Preference Optimization for Multimodal Large Language Models. *arXiv preprint arXiv:2406.11839*.
- Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. 2025. Grounded-VideoLLM: Sharpening Fine-grained Temporal Grounding in Video Large Language Models. In *EMNLP Findings*.
- Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024b. VideoHalluciner: Evaluating Intrinsic and Extrinsic Hallucinations in Large Video-Language Models. *arXiv preprint arXiv:2406.16338*.
- 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, and 20 others. 2025. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*.

- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NEX-T-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *CVPR*.
- Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: Mitigating Hallucination in Large Vision-Language Models via Vision-Guided Direct Preference Optimization. In *EMNLP Findings*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *ICCV*.
- Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. 2024a. EventHallusion: Diagnosing Event Hallucinations in Video LLMs. *arXiv preprint arXiv:2409.16597*.
- Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. [LLaVA-NeXT: A Strong Zero-shot Video Understanding Model](#).
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.

A Additional Implementation and Training Details

A.1 Data Generation

Detailed Data Generation Configuration. All stages of the counterfactual video generation pipeline are executed in a modular and parallelized manner, with keyframe extraction, action proposal, image editing, and video synthesis run independently across up to three NVIDIA A100 GPUs (40GB). Generating the $\sim 26k$ preference pairs in CounterVid required approximately 3,000 GPU-hours and about \$60 in LLM API cost for proposal, filtering, and refinement. Keyframe extraction uses SigLIP-SO400M¹ for caption-frame alignment, sampling frames with a temporal stride of 0.5s and averaging similarity over multiple spatial crops. Action proposal and filtering are performed with Claude-4-Sonnet using a temperature equal to 0.0 for deterministic outputs. End-frame generation is conducted with Qwen-Image-Edit² using the Lightning-4steps-V1.0 LoRA adapter³, running in FP8 precision with four diffusion steps, classifier-free guidance equal to 1.0, and denoising strength 0.9. An iterative refinement loop with a maximum of five attempts is used to ensure action correctness and scene consistency. Video synthesis employs Wan2.2-I2V-14B⁴ in FP8 precision with the I2V_lightx2v_4steps LoRA adapter⁵, generating videos at 680 $\times$ 384 resolution with 49 frames at 16 fps, corresponding to a duration of approximately 3.06 seconds.

Safety Filtering. We apply automated visual safety screening to the generated samples using NudeNet (Bedapudi, 2019), targeting sexual content, and Q16 (Schramowski et al., 2022), covering broader harmful categories such as gore and violence. These classifiers flag approximately 0.6% of generated samples as potentially unsafe; flagged samples are removed from the final released dataset. Because CounterVid is derived from PVD only through anchor frames, it does not include source metadata, and we inherit PVD’s privacy-protection procedures for the real anchors.

Human Annotation Protocol. Human evaluators use an interface that displays the visual input, the prompt, and the target answer for each sampled

¹[google/siglip-so400m-patch14-384](https://github.com/google/siglip-so400m-patch14-384)

²[Qwen/Qwen-Image-Edit](https://github.com/Qwen/Qwen-Image-Edit)

³[lightx2v/Qwen-Image-Lightning](https://github.com/lightx2v/Qwen-Image-Lightning)

⁴[Wan-AI/Wan2.2-I2V-A14B](https://github.com/Wan-AI/Wan2.2-I2V-A14B)

⁵[Comfy-Org/Wan_2.2_ComfyUI_Repackaged](https://github.com/Comfy-Org/Wan_2.2_ComfyUI_Repackaged)

preference item. For output-swap examples, the interface shows one visual input and two answers, and annotators evaluate whether the chosen answer is correct for that visual input. For input-swap examples, it shows the chosen and rejected visual inputs with a single answer, and annotators evaluate the answer only with respect to the chosen visual input. Separately, annotators rate the visual quality of the visual input, independently of answer correctness. This two-axis protocol prevents the “good” label from conflating semantic validity with visual fidelity and makes label noise and generation artifacts separately observable.

A.2 Training and Evaluation

Training Configuration. To train MixDPO, we fine-tune Qwen2.5-VL models (3B⁶ and 7B⁷) and InternVL3-9B⁸ using parameter-efficient adaptation. The vision encoder is frozen except for the vision-language merger module, while the language model is adapted with LoRA (Hu et al., 2022) applied to attention layers (rank 64, $\alpha=16$). Training is performed for three epochs at 2 fps with a learning rate of 1×10^{-6} using the AdamW optimizer (Loshchilov and Hutter, 2019), an effective batch size of 8 distributed over 8 GPUs, and a KL penalty coefficient of $\beta=0.7$. The reference policy π_{ref} is a frozen copy of the corresponding initial checkpoint.

Evaluation Configuration. For reproducibility, we apply a consistent pre-processing setup across all evaluation benchmarks. For all image and video inputs, we enforce a strict pixel budget where the maximum resolution is set to 151,200 pixels. Similarly, the minimum resolution is maintained at 100,352 pixels. Regarding temporal sampling, the model is limited to a maximum of 32 frames for all video-based benchmarks.

PaMi-VDPO Baseline Configuration. For the PaMi-VDPO baseline (Ding et al., 2025), we construct two augmentation-based rejected instances for each chosen video. Specifically, we apply *Crop*, which randomly crops at most 20% of the spatial region and resizes the result back to the original resolution, and *Shuffle*, which randomly permutes the temporal order of frames or short segments. The chosen video and both rejected instances form a single multi-instance training example. The PaMi-

Model	Temporal Ord.			Action Rec.			Avg
	FF	OL	BC	FF	MC	BC	
▼ Qwen2.5-VL-3B	29.8	1.6	48.9	28.0	48.4	64.6	36.9
PaMi-VDPO (PVD)	32.2	1.6	47.1	29.1	51.4	67.6	38.2
PaMi-VDPO (CtVid)	32.9	1.4	47.5	29.1	52.3	68.0	38.5
MixDPO (Ours)	31.0	16.7	64.0	29.5	60.0	70.3	45.2
▼ Qwen2.5-VL-7B	70.3	16.5	57.8	55.1	68.6	78.2	57.8
PaMi-VDPO (PVD)	42.3	13.2	57.6	33.0	64.2	74.1	47.4
PaMi-VDPO (CtVid)	42.2	13.4	57.6	32.9	64.4	74.5	47.5
MixDPO (Ours)	72.2	43.8	72.7	56.3	71.2	80.7	66.2
▼ InternVL3-9B	66.3	13.2	53.0	49.9	65.3	75.4	53.8
PaMi-VDPO (PVD)	65.4	12.8	56.0	50.4	64.4	75.8	54.1
PaMi-VDPO (CtVid)	66.5	14.9	52.3	53.7	65.3	74.3	54.5
MixDPO (Ours)	74.1	61.3	84.7	63.2	79.6	81.1	74.0

Table 6: **PaMi-VDPO variant breakdown on CounterVid.** We compare the PVD-anchor and CounterVid-anchor variants of the PaMi-VDPO baseline under the same augmentation-based negatives, and report the corresponding base and MixDPO results for context.

Samples	CLIP-Score	VQAScore
Real PVD pairs	0.268	0.720
Generated CounterVid pairs	0.251	0.673

Table 7: **Automatic semantic quality metrics.** We compare original PVD video-caption pairs and generated CounterVid pairs under the same CLIP-Score (Hessel et al., 2021) and VQAScore (Lin et al., 2024b) protocol.

VDPO objective derives prompt/model-dependent weights from token-level divergence scores and jointly applies them to the two negative rewards. Both variants use the same three-epoch budget and backbone-specific optimization settings as the matched comparison. In the main tables, the chosen videos are CounterVid-generated clips; Appendix B reports both the CounterVid- and PVD-anchor variants.

B Additional Results and Analyses

Detailed Comparison with PaMi-VDPO. In Table 6, we separate two controlled variants that use the same prompt-aware multi-instance objective and augmentation-based rejected instances, but differ in the source of the chosen videos. In **PaMi-VDPO (PVD)**, the chosen videos are original PVD clips; in **PaMi-VDPO (CtVid)**, they are CounterVid-generated clips, matching the chosen-video distribution used by MixDPO in the main paper. The CtVid variant therefore controls for the chosen-video source while replacing our semantic counterfactual negatives with generic Crop and Shuffle perturbations. In both variants, the two negatives are processed jointly rather than as independent DPO pairs.

⁶Qwen/Qwen2.5-VL-3B-Instruct

⁷Qwen/Qwen2.5-VL-7B-Instruct

⁸OpenGVLab/InternVL3-9B

Model	Setting	Temporal Ord.			Action Rec.			Avg.
		FF	OL	BC	FF	MC	BC	
Qwen2.5-VL-7B	Default: $\lambda=1, \beta=0.7, 70/30$	72.2	43.8	72.7	56.3	71.2	80.7	66.2
	$\lambda=0.5$	72.2	43.8	72.7	56.2	70.5	80.4	66.0
	$\lambda=2.0$	72.3	42.2	72.5	56.2	71.0	80.2	65.7
	$\beta=0.5$	72.4	51.1	77.4	56.6	69.0	81.5	68.0
	$\beta=1.0$	71.6	35.4	68.2	56.0	71.0	80.4	63.8
	v/t = 50/50	72.3	42.8	72.9	56.3	71.0	80.7	66.0
	v/t = 90/10	72.1	46.4	74.1	56.3	70.1	80.0	66.5
InternVL3-9B	Default: $\lambda=1, \beta=0.7, 70/30$	74.1	61.3	84.7	63.2	79.6	81.1	74.0
	$\lambda=0.5$	74.2	61.1	84.1	63.0	79.8	81.9	74.0
	$\lambda=2.0$	74.0	60.5	84.5	63.1	79.3	81.3	73.8
	$\beta=0.5$	74.8	62.5	84.9	63.5	79.6	82.1	74.6
	$\beta=1.0$	73.0	58.0	82.1	62.7	79.8	80.9	72.7
	v/t = 50/50	73.7	59.5	82.7	62.6	79.6	82.1	73.4
	v/t = 90/10	74.6	61.9	85.7	62.9	79.3	81.9	74.4

Table 8: **Hyperparameter and preference-composition sensitivity on CounterVid.** Each block reports a matched default configuration and changes only the indicated factor. 70/30 denotes the visual/textual preference ratio.

Cross-Model Agreement. To assess dependence on a single closed-source evaluator, we evaluate the two Claude-based filtering stages (action proposal filtering and edited-frame evaluation) on 5,000 samples using two open-source VLM judges, Qwen2.5-VL-7B and InternVL3-14B. The accept/reject decisions agree with Claude in 93.6% of cases for Qwen2.5-VL-7B and 76.3% for InternVL3-14B. This suggests that the filtering criteria are not unique to Claude. It does not, however, establish full interchangeability of the image editor or video generator, which we leave for future cross-generator validation.

Automatic Semantic Quality Metrics. Table 7 complements the human study shown in Table 1 with automatic semantic metrics. Specifically, CLIP-Score (Hessel et al., 2021) measures cross-modal video-caption alignment, while VQAScore (Lin et al., 2024b) estimates answer consistency through visual question answering. Under the same evaluation protocol, generated CounterVid pairs remain close to the original PVD pairs (0.251 vs. 0.268 CLIP-Score and 0.673 vs. 0.720 VQAScore), suggesting that the pipeline largely preserves semantic alignment while introducing controlled action or temporal variations.

B.1 Hyperparameter Sensitivity

We vary one factor at a time around the paper configuration, using the same source pool, train/validation partition, total number of examples, and CounterVid evaluation protocol. Table 8 reports the six task submetrics for the visual-preference loss weight λ , the visual/textual preference mixture, and the DPO coefficient β .

The λ sweep confirms stability across a four-

Model	Temporal Ord.			Action Rec.			Avg
	FF	OL	BC	FF	MC	BC	
▼ InternVL3-9B	66.3	13.2	53.0	49.9	65.3	75.4	53.8
T-pref-DPO	72.5	56.4	80.9	62.2	78.5	81.5	72.0
MixDPO (Ours)	74.1	61.3	84.7	63.2	79.6	81.1	74.0
▼ InternVL3-14B	68.3	15.9	63.3	54.4	66.8	72.7	56.9
T-pref-DPO	76.8	66.0	91.9	64.7	82.2	82.7	77.4
MixDPO (Ours)	77.6	66.4	92.3	65.5	82.6	83.5	78.0

Table 9: **Scaling experiment on CounterVid.** Accuracy (%) for InternVL3-9B and InternVL3-14B under the matched Base, textual-preference DPO, and MixDPO configurations. FF: free-form; OL: order list; BC: binary choice; MC: multiple choice.

fold range: average accuracy changes by at most 0.5pp on Qwen2.5-VL-7B and 0.2pp on InternVL3-9B. Preference composition is similarly stable. Although 90/10 improves the average by 0.3pp and 0.4pp, respectively, the gains are not consistent across submetrics; we therefore retain the 70/30 setting rather than selecting for a marginal advantage on this benchmark. The β sweep has a larger effect: $\beta=0.5$ improves the average by 1.8pp and 0.6pp, whereas $\beta=1.0$ lowers it by 2.4pp and 1.3pp. Because these variants were evaluated only on CounterVid, we retain $\beta=0.7$ as a conservative default.

B.2 Scaling to InternVL3-14B

To examine whether the behavior observed in the main experiments extends to a larger model, we train InternVL3-14B using the same CounterVid data, MixDPO objective, and training protocol. Table 9 reports this experiment together with InternVL3-9B as a matched reference, comparing only Base, T-pref-DPO, and MixDPO on CounterVid. At 14B, MixDPO improves over Base by 21.1 points on average and over T-pref-DPO by 0.6

Failure category	Count	Errors (%)
Failed action realization	39	84.8
Generation artifacts	18	39.1
Ambiguous action label	11	23.9
Scene identity drift	0	0.0
Temporal concatenation artifacts	0	0.0

Table 10: **Failure-mode breakdown for the 46 samples flagged during human evaluation.** Flags indicate an incorrect or ambiguous label, poor visual quality, or both; multiple failure categories may apply per sample.

points, with gains on all six submetrics. This provides evidence that visual-preference supervision remains useful at this scale, while not establishing behavior at frontier-scale models.

B.3 Synthetic-Data Failure Analysis

Following the human study in Sec. 4.2, we categorize the 46 unique samples flagged for an incorrect or ambiguous label, poor visual quality, or both. Table 10 reports the resulting failure modes. Categories are not mutually exclusive, so counts and percentages need not sum to 46 or 100%. Failed action realization is the dominant residual issue, followed by visible generation artifacts and ambiguous action labels. We observe no scene-identity drift or temporal-concatenation failures in this set. Fig. 5 complements this aggregate analysis with a successful generation and representative cases from all three nonzero failure categories. The examples distinguish clean action realization from an unchanged intended state, visible synthesis inconsistencies, and action descriptions that are not visually separable.

C Evaluation Benchmarks Details

To provide a comprehensive evaluation of model performance, we employ a diverse set of benchmarks, targeting both specific hallucination modes and general video understanding capabilities.

C.1 Video Hallucination Benchmarks

We evaluate on four benchmarks designed to assess complementary aspects of multimodal hallucination, ranging from object existence to event-level and temporal reasoning.

VideoHalluciner (Wang et al., 2024b). It comprises approximately 1,000 samples evaluated via binary-choice question answering. It distinguishes between intrinsic hallucinations, corresponding to incorrect descriptions of content present in the video (*e.g.*, object relations, temporal sequencing, seman-

tic details), and extrinsic hallucinations, which involve fabricating content absent from the video, often driven by language priors, including factual errors and unsupported inventions.

VidHalluc (Li et al., 2025a). VidHalluc is a large-scale benchmark containing approximately 5,000 videos and 9,000 question-answer pairs. It evaluates models across binary-choice, multiple-choice, sorting, and open-ended formats, targeting action hallucination (misidentification of actions), temporal sequence hallucination (errors in event ordering), and scene transition hallucination (misinterpretation of changes across scenes). In our experiments, we report results on the temporal sequence hallucination (TSH) samples.

VideoHallu (Li et al., 2025b). This benchmark consists of approximately 3,000 synthetic videos designed for short-answer question answering. It evaluates hallucinations across alignment (entity counting, recognition, and entity properties), spatial-temporal consistency (camera, spatial, and temporal dynamics), physics reasoning (constraints, properties, motion, and state transitions), and commonsense (general knowledge and reasoning capabilities).

EventHallusion (Zhang et al., 2024a). EventHallusion comprises 400 videos and about 700 question-answer pairs, focusing on event-level hallucinations driven by parametric knowledge (language priors). It categorizes videos into three types: *entire rare events*, which depict uncommon events to test reliance on priors; *interleave*, which features a mix of common and rare events occurring sequentially; and *misleading*, which pairs videos of common events with customized questions designed to trigger misleading event demonstrations.

C.2 General Video Understanding Benchmarks

To ensure that mitigating hallucinations does not compromise general capabilities, we also evaluate on three standard benchmarks.

VideoMME (Fu et al., 2025). VideoMME is a comprehensive benchmark for multimodal video understanding that assesses the model’s ability to process and reason over diverse video domains.

NExT-QA (Xiao et al., 2021). This benchmark is specifically designed to evaluate causal and temporal action reasoning. It challenges models to explain “why” and “how” actions occur, moving beyond simple descriptive tasks.

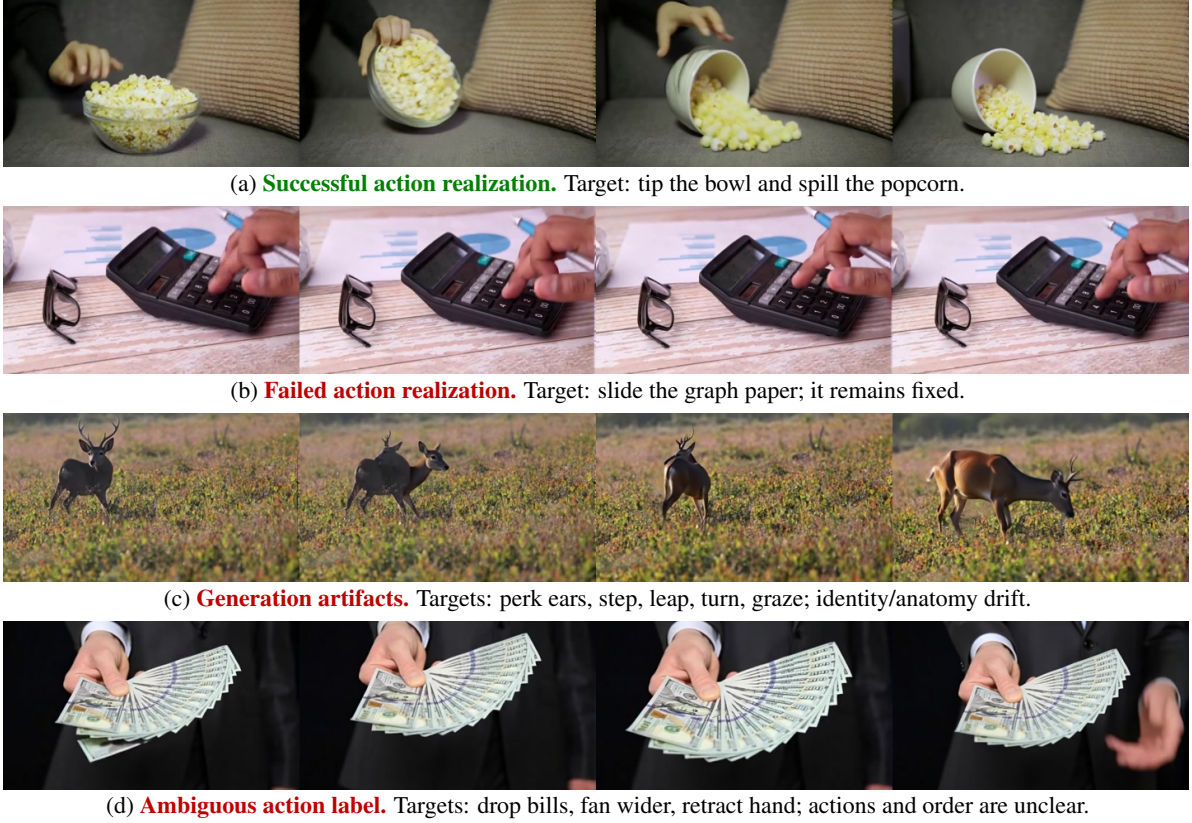


Figure 5: **Representative successful generation and failure cases.** Each row shows four uniformly spaced frames from one generated clip or concatenated action sequence. The examples make action realization and residual synthesis errors directly inspectable; the aggregate failure distribution is reported in Table 10.

TempCompass (Liu et al., 2024d). Focused on temporal commonsense reasoning, it employs multiple-choice, binary-choice, and caption matching formats to test understanding of temporal concepts such as duration, frequency, and event order.

D Generation Pipeline Prompts

Our counterfactual generation pipeline is implemented through a sequence of structured prompts that govern action proposal, action filtering, image editing, and iterative refinement. For completeness and reproducibility, we report the system prompts used in each stage of the pipeline.

Action Proposal and Filtering. Conditioned on the anchor frame I^{start} and the original caption C^{real} , a multimodal LLM is first prompted to propose N alternative action captions $\{C_i\}_{i=1}^N$. As shown in Fig. 6, the action proposal prompt enforces that each action is contextually appropriate, involves visible subjects or objects, is achievable via image editing, and is semantically distinct from the others. To remove implausible or redundant proposals, a second prompt evaluates each action according to subject presence, physical and contextual feasibility, and semantic uniqueness (Fig. 7).

Actions that fail any criterion (such as relying on absent objects or being minor variants of another action) are discarded, ensuring a diverse set of visually expressible hard negatives.

Editing Instruction Generation and Refinement.

For each retained action caption C_i , we prompt the same model to generate a structured image-editing instruction that transforms the anchor frame into a plausible final-state depiction of the action. As illustrated in Fig. 8, this prompt constrains the instruction to preserve scene identity, camera parameters, and unrelated content, while explicitly describing the visible outcome of the action.

Because image editing may introduce semantic drift or artifacts, we employ an iterative verification loop. After each edit, an evaluation prompt compares the original and edited frames to determine whether the intended action has been correctly realized while maintaining scene consistency (Fig. 9). If the edit fails, a refinement prompt generates a new instruction explicitly differentiated from all previous attempts and targets the observed failure modes (Fig. 10). This refinement process is repeated up to a fixed number of attempts, yielding a validated end frame for subsequent video synthesis.

You are given a single video frame (image) and the original video caption.

ORIGINAL CONTEXT (caption): {caption}

TASK: Propose EXACTLY {num_actions} simple, realistic actions that could plausibly happen in THIS scene. Each action must be FUNDAMENTALLY DIFFERENT from all others.

REQUIREMENTS FOR EACH ACTION:

- Must be contextually appropriate for what you see in the frame
- Must involve subjects/objects actually visible in the frame
- Must be achievable through image editing (transforming this frame to show the final post-action state)
- Must be UNIQUE – no two actions should be similar or variations of each other
- Keep actions simple – avoid complex multi-step actions

UNIQUENESS REQUIREMENTS:

- Each action must involve DIFFERENT types of movement, interaction, or change
- Avoid similar actions like “wave with right hand” vs. “wave with left hand”
- Avoid variations of the same action like “open door” vs. “open door wider”
- Focus on completely different action categories: gestures, movements, interactions, expressions, object manipulations, etc.

EXAMPLES OF GOOD DIVERSE ACTIONS:

- Person waves hand (gesture)
- Person turns head to look left (movement)
- Person picks up object (interaction)
- Door opens (object change)
- Person smiles (expression)

EXAMPLES OF BAD SIMILAR ACTIONS (avoid these patterns):

- Person waves right hand + Person waves left hand
- Person opens door + Person closes door
- Person looks left + Person looks right

FOR EACH ACTION, PROVIDE:
action_caption: Clear, concise description of what’s happening in the action

OUTPUT FORMAT (JSON only, no extra text):

```
{
  "actions": [
    {"action_id": 0, "action_caption": "..."},
    {"action_id": 1, "action_caption": "..."},
    {"action_id": 2, "action_caption": "..."},
    ...
  ]
}
```

Return ONLY the JSON.

Figure 6: **Action proposal prompt.** The VLM is instructed to generate N semantically distinct and physically plausible actions conditioned on the anchor frame, which serve as candidate counterfactuals.

You are evaluating proposed actions for a video frame to determine if they are realistic, feasible, and UNIQUE from each other.

ORIGINAL VIDEO CONTEXT: {caption}

PROPOSED ACTIONS: {actions}

EVALUATION CRITERIA:

1) Subject Presence: Are all subjects/objects mentioned in the action actually visible in the frame?

2) Physical Feasibility: Is the proposed action physically possible given the current frame?

3) Contextual Appropriateness: Does the action make sense in this scene/environment?

4) General Feasibility: Could this action plausibly happen in this scene?

5) UNIQUENESS: Is this action fundamentally different from all other proposed actions?

UNIQUENESS EVALUATION:

- Actions are TOO SIMILAR if they involve the same type of movement/gesture (e.g., “wave right hand” vs. “wave left hand”)

- Actions are TOO SIMILAR if they are variations of the same action (e.g., “open door” vs. “close door”)

- Actions are TOO SIMILAR if they involve the same body part doing similar movements (e.g., “turn head left” vs. “turn head right”)

- Actions PASS uniqueness if they involve completely different types of actions (gesture vs. expression vs. object interaction vs. movement)

YOUR TASK: For each action, determine if it passes ALL evaluation criteria. An action FAILS if:

- It mentions subjects/objects not visible in the frame
- It's physically impossible given the current pose/position
- It's completely inappropriate for the scene context
- It's unrealistic for this type of scene
- **It's too similar to another proposed action**

RESPONSE FORMAT: Provide your response in this exact JSON format:

```
{
  "evaluations": [
    {
      "action_id": 0,
      "passed": True/False,
      "uniqueness_check": "Explanation of uniqueness evaluation",
      "similarity_issues": "List any actions this is too similar to, or None"
    },
    ...
  ],
  "overall_assessment": "Summary of the evaluation results including uniqueness analysis",
  "uniqueness_summary": "Overall assessment of how unique the actions are from each other"
}
```

Analyze the frame and evaluate each proposed action for both realism and uniqueness.

Figure 7: **Action quality evaluation prompt.** This prompt filters proposed actions by enforcing subject presence, physical and contextual feasibility, and semantic uniqueness within the scene.

ROLE: You are an expert image-editing prompt engineer.

GOAL: Look at this INITIAL_FRAME (image) and an ACTION_CAPTION (text describing what has already happened by the final frame), produce a single, executable edit instruction for an image-editing model that transforms the initial frame into the believable final state, while preserving all unrelated details for video continuity (FFLF).

INPUTS:

- INITIAL_FRAME: attached the image of the starting frame
- ACTION_CAPTION: a short description of the event's outcome by the final frame

OUTPUT FORMAT:

Return ONLY the following JSON:

```
{“edit_prompt”: “one concise, model-ready edit instruction”}
```

HOW TO WRITE edit_prompt:

- 1) Start with a precise verb: ADD / REMOVE / REPLACE / MODIFY.
- 2) Describe the visible final-state evidence of the action (effects on objects, materials, surfaces).
- 3) Constrain for continuity:
 - Keep everything else unchanged.
 - Preserve camera, composition, perspective, framing, time of day, grain/noise.
 - Preserve subject identity (faces, clothing, hairstyle) unless explicitly edited.
 - Preserve existing text layout (font, size, alignment, perspective) unless explicitly edited.
 - No camera move, no focal length or depth-of-field change.
- 4) Ensure physical plausibility:
 - Match lighting direction, shadow softness, reflections, occlusions, contact shadows.
 - Respect materials (wetness, scorch, dents, debris, splash, smoke) consistent with the scene.
 - Add subtle motion evidence only if warranted, without altering pose/composition.
- 5) Pinpoint WHERE the change occurs using anchors from the image (e.g., “on the desk in front of the mug”, “upper-left window pane”, “subject’s right hand”).
- 6) Include a brief negative clause to prevent artifacts:
 - “no extra people/limbs, no duplicates, no distortion, no style drift, no warped text”.

STYLE & BREVITY:

- 1-3 sentences total.
- Be conservative; prefer minimal edits needed to depict the outcome.
- Do not ask questions; make your best, plausible inference from inputs.
- Output only the JSON object above—no explanations.

FEW-SHOT EXAMPLES (follow style, not content):

Example A

INITIAL_FRAME: Man holding a ceramic mug on a wooden desk; window light from left.

ACTION_CAPTION: He spills coffee from the mug onto the desk.

Expected: {“edit_prompt”: “ADD a fresh coffee spill spreading from the mug’s rim onto the wooden desktop with dark wet sheen, small droplets, and a thin run toward the front edge; match left-window lighting for highlights and contact shadows; preserve man, pose, mug, camera, perspective, grain, and desk items. Keep everything else unchanged; no extra hands, no new objects, no distortion.”}

Example B

INITIAL_FRAME: Storefront banner reading “SUMMER SALE”, angled perspective.

ACTION_CAPTION: The sale ended; banner now says ‘WELCOME BACK’.

Expected: {“edit_prompt”: “REPLACE the banner text with ‘WELCOME BACK’, matching the original font, size, color, material, weathering, folds, and perspective warp; keep stitching and fabric texture consistent. Keep everything else unchanged; no warped letters, no background shift, no added objects.”}

Example C

INITIAL_FRAME: Woman jogging in a park, ponytail; late-afternoon sun from right.

ACTION_CAPTION: She finishes in light rain; she’s slightly wet and breathing.

Expected: {“edit_prompt”: “MODIFY the scene to add a light drizzle and subtle wetness on hair and jacket with fine highlights; add faint condensed breath near the mouth; keep sun direction from right and existing shadowing intact. Keep everything else unchanged; no extra people, no heavy downpour, no camera change.”}

CALL TEMPLATE (fill and run):

Use the instructions above.

INITIAL_FRAME: attached image frame

ACTION_CAPTION: {action_caption}

Return only: {“edit_prompt”: “your concise, model-ready edit instruction”}

Figure 8: **Editing instruction generation prompt.** The high-level action caption is converted into structured image-editing instructions suitable for the image editing model.

Look at these two images: the original image and the edited image.
EDITING INSTRUCTION: {editing_prompt}
 Compare the original and edited images. Has the editing instruction been correctly applied in the edited image?
 First, provide your evaluation as either “YES” or “NO”.
 - “YES” if the editing instruction has been correctly applied.
 - “NO” if it has not been correctly applied or if the changes are incorrect.
 If your evaluation is “NO”, then provide a short explanation of what is wrong with the generation.
OUTPUT FORMAT:
 EVALUATION: YES/NO
 EXPLANATION: Only provide if evaluation is NO – short explanation of what went wrong

Figure 9: **Edited frame evaluation prompt.** The prompt compares the original and edited frames to verify that the intended action is correctly realized while preserving scene identity.

ROLE: You are an expert image-editing prompt engineer specialized in Qwen-style edit instructions.
SITUATION: Multiple attempts have failed. Produce ONE new, executable edit instruction that is fundamentally different from all prior attempts.
INPUTS:
 - DESIRED OUTCOME: {desired_outcome}
 - CURRENT FAILED PROMPT: {original_prompt}
 - ALL PREVIOUS FAILED PROMPTS: {failed_list}{failure_block}
PLAYBOOK ADHERENCE (apply silently; do NOT output this section):
 - Choose exactly ONE editing family: Text edit / Local appearance / Global restyle (semantic) / Micro or region edit / Identity control / Poster & composite / Lighting & camera.
 - Include the continuity clause: “Keep everything else unchanged”.
 - If editing text, preserve font, size, color, alignment, perspective.
 - Preserve subject identity (faces, clothing, hairstyle) unless the outcome requires a change.
 - Match scene physics: lighting direction, shadow softness, reflections, materials, occlusions, perspective.
 - Add a concise negative clause (e.g., “no duplicates, no distortion, no warped text, no style drift”).
 - Pinpoint the location with unambiguous anchors visible in the image.
DIFFERENTIATION MANDATE (must differ on ≥ 2 axes from ALL failures):
 1) VERB FAMILY: switch to a new one (ADD / REMOVE / REPLACE / MODIFY) not used before.
 2) FOCUS: change target (subject $\longleftrightarrow$ environment/background $\longleftrightarrow$ lighting/shadows/reflections $\longleftrightarrow$ layout/spacing).
 3) STYLE: change descriptive style (hyper-specific anchors/measurements $\longleftrightarrow$ conceptual brevity; technical terms $\longleftrightarrow$ plain language).
 4) SPATIAL ANCHORING: use new landmarks; switch relative $\longleftrightarrow$ absolute positioning.
 5) PHRASEOLOGY: avoid reusing distinctive phrases or any 3-word sequence from failed prompts.
OUTPUT RULES (hard constraints):
 - Start with EXACTLY one verb: ADD, REMOVE, REPLACE, or MODIFY.
 - 1-2 sentences total, ≤ 55 words.
 - Forbid camera/pose/focal changes unless explicitly required by DESIRED OUTCOME.
 - Depict only visible final-state evidence; prefer the minimal change that unambiguously communicates the outcome.
 - No meta-commentary, no lists, no JSON, no quotes, no hedging.
RETURN Only the new editing instruction as plain text, in 1-2 sentences (≤ 55 words), starting with the chosen verb.

Figure 10: **Refinement prompt.** When an edit attempt fails, this prompt generates a new image-editing instruction that is explicitly differentiated from prior attempts and targets previously observed failure modes.