VIDEOSTF: Stress-Testing Output Repetition in Video Large Language Models

Yuxin Cao[1] **Wei Song**[2*] **Shangzhi Xu**[3] **Jingling Xue**[3] **Jin Song Dong**[1*]

[1]*National University of Singapore, Singapore*
[2]*Griffith University, Australia*
[3]*University of New South Wales, Australia*

Abstract

Video Large Language Models (VideoLLMs) have achieved strong performance on video understanding tasks, yet existing benchmarks evaluate only *what* models predict, leaving the stability of *how* they generate largely unexamined. We surface a previously underexplored generation failure of VideoLLMs, defined as *output repetition*, in which the decoder collapses into self-reinforcing loops of repeated phrases or sentences, and present VIDEOSTF, a benchmarking framework for systematically measuring, stress-testing, and exploiting this failure mode. VIDEOSTF formalizes repetition with three complementary n-gram-based metrics, ships a standardized testbed of 10,000 diverse videos, and provides a library of controlled temporal stressors. Across 10 advanced VideoLLMs, VIDEOSTF reveals four key findings: (i) repetition is pervasive on unperturbed videos and stable across commonly used frame counts, with repetition rates up to 91%; (ii) it spans a severity spectrum from mild redundancy to token-cap loops, and is highly amplified by temporal perturbations; (iii) temporal stressors form a practical black-box attack surface, flipping benign videos into repetitive ones with tens of queries and high attack success rates (up to 98%), and (iv) repetition is not explained by visual redundancy, its amplification tracks local temporal disruption, and only repetition penalties reduce it among common mitigations such as top-k sampling, input filtering, and prompt variation, but increasing the penalty weakens visual grounding. VIDEOSTF reframes generation stability as a useful and complementary evaluation axis for VideoLLMs and provides the tools to study it. The project page is available at `https://videostf.github.io/`.

1 Introduction

Video has become a pervasive modality in the modern digital era, and the ability of artificial intelligence systems to understand videos has reached a new peak with the emergence of Video Large Language Models (VideoLLMs) [1–3]. Benefiting from their strong temporal reasoning capability and powerful multimodal representation learning, VideoLLMs have been widely adopted in video understanding tasks such as video captioning [4, 5] and question answering [6, 7].

However, we identify **output repetition**, a severe autoregressive generation failure in VideoLLMs arising from interactions between temporal visual inputs and language generation. As shown in Figure 1, a typical normal output produces text that accurately and coherently describes a video, whereas a repetitive output degenerates into repeated phrases or sentences, which can degrade response quality and user experience. In extreme cases, such repetition persists until the model reaches its output token limit, resulting in unnecessary computation and enabling adversarial exploitation [8],

*Corresponding authors: Wei Song (w.song@griffith.edu.au) and Jin Song Dong (dcsdjs@nus.edu.sg).

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

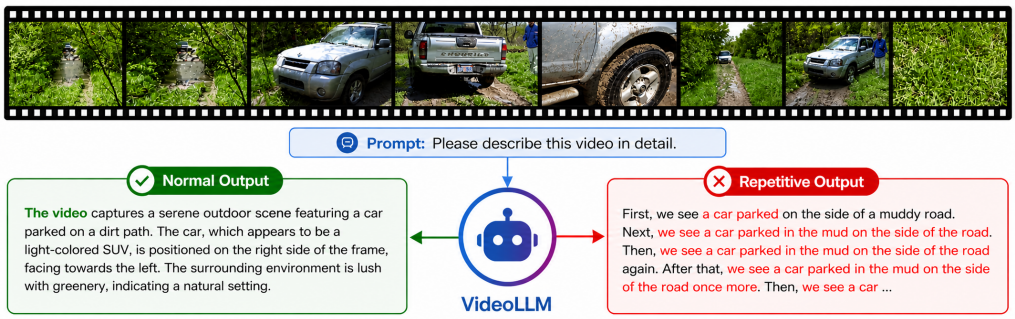

Figure 1: Examples of normal and repetitive outputs.

such as denial-of-service, undermining the reliability of VideoLLMs. While numerous benchmarks have been conducted for VideoLLMs [9–16], they primarily focus on task correctness, factual errors, and hallucinations, overlooking generation repetition.

In this work, we systematically study this generation failure in VideoLLMs and introduce VIDEOSTF, a benchmark for measuring, probing, and exploiting output repetition. VIDEOSTF first formalizes output repetition using three n-gram-based metrics: a repetition rate that captures whether repetition occurs, a repetition intensity that quantifies the extent of duplicated patterns, and an entropy-based metric that quantifies repetition through lexical diversity. Furthermore, VIDEOSTF provides a standardized video testbed of 10,000 examples, with a temporal stressor library that defines a set of temporal transformations for stress-testing output repetition in VideoLLMs.

Using VIDEOSTF, we conduct three tests: pervasive testing to characterize repetition under natural inputs, temporal stress testing to probe sensitivity to temporal transformations, and adversarial exploitation to assess whether repetition can be actively induced as a black-box attack. We evaluate 10 VideoLLMs, including seven LLM-centric models, LLaVA-Video-7B-Qwen2, LLaVA-Video-7B-Qwen2-Video-Only [17], LLaVA-NeXT-Video-7B, LLaVA-NeXT-Video-7B-DPO, LLaVA-NeXT-Video-32B-Qwen [18], VideoLLaMA2 [19], and ShareGPT4Video [5], as well as three native multimodal models, InternVL3.5-8B [3], Qwen3-VL-8B-Instruct [20], and Molmo2-8B [21].

Pervasive Testing. Using the standardized testbed, VIDEOSTF benchmarks each VideoLLM under varying numbers of sampled frames, treating temporal granularity as a controlled variable. Across all 10 models, repetition on unperturbed videos is widespread and stays stable across commonly used frame counts, indicating a pervasive failure mode that persists under the frame settings typical of inference. The repetition typically involves recurring references to the same people, objects, or scenes. Moreover, characteristic looping expressions such as "continues to" are prevalent in the video captions used to train some of these models, suggesting a model-internal tendency to enter self-reinforcing repetition patterns.

Temporal Stress Testing. If repetition stems from how VideoLLMs aggregate temporally similar visual cues, perturbing temporal structure while preserving semantics should sharpen the effect. We therefore apply five stressor families (Add, Delete, Replace, Reverse, and Shuffle) and compare each model's outputs to its original one. Repetition rates rise under most stressors, reaching over 90% in the most affected cases, indicating a substantial sensitivity to structured temporal perturbations.

Adversarial Exploitation. If a single temporal stressor amplifies repetition population-wide, the same stressor may also *flip* individual benign videos into repetitive ones. We therefore cast the stressor library as a query-efficient black-box attack: under a realistic threat model in which an attacker can modify the sampled frames used for inference but has no access to model internals, attacks succeed with tens of queries (attack success rate up to 98%). Output repetition is therefore not merely a diagnostic artifact *measured* by VIDEOSTF but an exploitable failure mode of modern VideoLLMs.

Beyond the three benchmarking tests, VIDEOSTF further probes the *drivers* of repetition and uncovers four key findings: (i) repetition spans a severity *spectrum* and lowers the information density of descriptions; (ii) it is not explained by visual redundancy, since neither highly redundant inputs nor redundancy filtering raise or remove it; (iii) *local* temporal disruption is the feature-level signal that tracks its amplification; and (iv) a training-data audit links repetition to video-caption supervision, while among common mitigation strategies, only increasing the repetition penalty reduces repetition,

albeit at the cost of weaker visual grounding. These findings extend VIDEOSTF from a measurement tool to a diagnostic framework that points toward components that future work needs to revisit.

In summary, we make the following contributions:

- We identify output repetition as an underexplored generation-stability failure in VideoLLMs and introduce VIDEOSTF, a benchmarking framework designed to expose and measure this failure through three complementary n-gram-based metrics, a standardized testbed of 10,000 diverse videos, and a library of controlled temporal stressors.
- Using VIDEOSTF, we benchmark 10 representative VideoLLMs and show that repetition is pervasive on unperturbed videos (repetition rate up to 91%), stable across commonly used frame counts, and strongly amplified by temporal perturbations even when semantic content is preserved.
- We demonstrate that the same temporal stressors constitute a practical black-box attack surface, flipping benign videos into repetitive ones with tens of queries (attack success rate up to 98%), elevating output repetition from a quality issue to a reliability and security concern.
- Beyond measurement, controlled analyses show that repetition is not explained by visual redundancy, and that Δrepetition correlates with Δadjacent-frame similarity alone. We further trace the looping style to video-caption training data and quantify the grounding cost of repetition penalties.

2 Related Work

VideoLLMs. VideoLLMs [3, 5, 17–19] extend LLMs to video understanding by conditioning language generation on temporally ordered visual inputs. They support tasks such as video captioning, video question answering, and video summarization. Existing VideoLLMs broadly fall into two paradigms. LLM-centric models retain a pretrained LLM as the generator and incorporate video through visual encoders and temporal projection modules [17–19, 22–25]. In contrast, native multimodal models such as InternVL3.5-8B [3] and Qwen3-VL-8B-Instruct [20] adopt unified multimodal backbones for jointly modeling vision and language. Despite architectural differences, both paradigms implicitly assume stable autoregressive generation once visual understanding succeeds.

VideoLLM Benchmarks. Existing VideoLLMs benchmarks evaluate video understanding accuracy [9–11], temporal reasoning [12], and semantic reliability, including factual errors [13], safety [14], hallucinations [15], and behavioral biases such as sycophancy [16]. These benchmarks primarily assess *what* a model predicts, while largely overlooking *how* generation unfolds. In particular, the stability of autoregressive generation is rarely examined, leaving output repetition—a severe degeneration phenomenon in VideoLLMs—largely unmeasured and unaccounted.

Output Repetition. Output repetition has been increasingly studied in text-only LLMs, where it has been attributed to intrinsic model distributions [26], repetitive patterns in training data [27], and identifiable internal activation features [28]. These studies establish repetition as a fundamental challenge in autoregressive generation that cannot be fully resolved by decoding strategies alone, even under deterministic decoding. However, they do not consider video-conditioned generation, where repetition can arise from interactions between temporal visual inputs and language decoding. In contrast to prior work, we study output repetition as a video-conditioned generation failure and introduce a dedicated framework for measuring and stress-testing this phenomenon in VideoLLMs.

3 VIDEOSTF

3.1 Output Repetition Formulation

A VideoLLM takes a video $\mathcal{V}$ as input, consisting of T sampled frames: $\mathcal{V} = \{I_1, \ldots, I_T\}$. Each frame I_t is encoded by a visual encoder ϕ_v to produce frame-level features $\mathbf{Z}_t = \phi_v(I_t)$, which are then aggregated into video-level features $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_T\}$. Then, a projector ϕ_p maps $\mathbf{Z}$ into the language model embedding space to produce projected visual tokens $\mathbf{H} = \phi_p(\mathbf{Z})$. The projected visual tokens $\mathbf{H}$ are subsequently concatenated with prompt embeddings $\mathbf{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_L\}$, where L denotes the token length of the prompt, to form a multimodal input sequence. Finally, the underlying LLM $\mathcal{F}$ generates the textual output conditioned on the fused sequence: $\mathbf{Y} = \mathcal{F}(\mathbf{H}; \mathbf{Q})$. Given a video input, the generated text output $\mathbf{Y}$ can be decomposed into M token-level units, denoted as $\mathbf{Y} = \{y_1, \ldots, y_M\}$, where each y_i denotes a single token. To formalize repetition, we employ three metrics based on n-gram [29], capturing repetition occurrence, intensity, and severity, respectively.

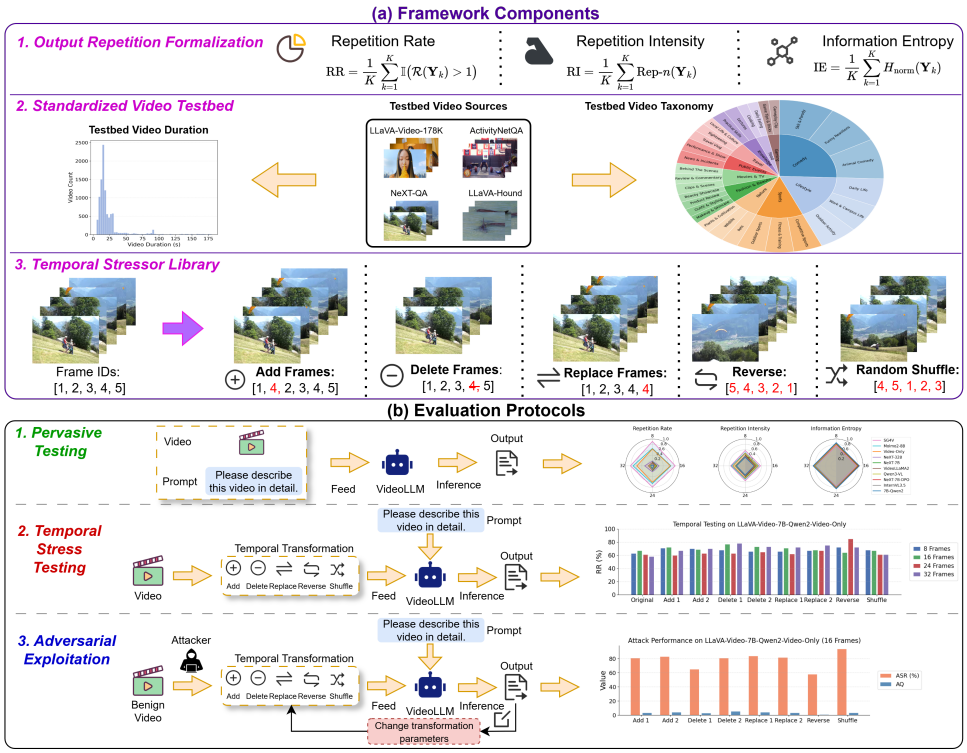

Figure 2: VIDEOSTF. (a) *Framework Components*. VIDEOSTF comprises three n-gram-based repetition metrics, a standardized testbed of 10,000 videos with diverse durations and content categories, and a library of controlled temporal stressors for applying temporal transformations. (b) *Evaluation Protocols*. VIDEOSTF assesses output repetition through three tests. *Pervasive Testing* reveals widespread repetition across VideoLLMs under different frame sampling settings across all three metrics. *Temporal Stress Testing* shows that temporal transformations amplify repetition. *Adversarial Exploitation* demonstrates that, through temporal transformations, videos with normal outputs can be efficiently induced to become repetitive with high success rates and few queries.

Repetition Rate (RR). We first quantify whether an output exhibits repetition using an n-gram-based [26, 28, 29] repetition rate. Specifically, given an output $\mathbf{Y}$, an n-gram is defined as any contiguous subsequence $(y_i, y_{i+1}, \ldots, y_{i+n-1})$ of length n in $\mathbf{Y}$. Let $\mathcal{S}(\mathbf{Y})$ denote the multiset of all contiguous n-grams in $\mathbf{Y}$, and $\mathrm{count}(s)$ denote the number of occurrences of an n-gram $s \in \mathcal{S}(\mathbf{Y})$, we can define the maximum number of times any n-gram is repeated in the output $\mathbf{Y}$ as: $\mathcal{R}(\mathbf{Y}) = \max_{s\in\mathcal{S}(\mathbf{Y})} \mathrm{count}(s)$, Given a set of K generated outputs $\{\mathbf{Y}_k\}_{k=1}^{K}$, the dataset-level RR is defined as:

$$\mathrm{RR} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}\big(\mathcal{R}(\mathbf{Y}_k) > \tau\big), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. We set $\tau = 1$ in our main experiments to capture the most direct repetition, and provide further analysis on the larger τ (more severe repetition) in Section 5. Thus, RR measures the fraction of outputs that contain at least one repeated n-gram.

Repetition Intensity (RI). While RR measures whether repetition occurs, it does not capture the extent of duplication within an output. We measure repetition intensity based on Rep-n metric [26, 30], which quantifies the proportion of duplicated n-grams:

$$\text{Rep-}n(\mathbf{Y}) = 1 - \frac{|\text{unique } n\text{-grams}(\mathbf{Y}, n)|}{M - n + 1}. \tag{2}$$

where $M - n + 1$ denotes the total number of contiguous n-grams in the sequence. We assess RI at the dataset-level by averaging Rep-n over all K outputs:

$$\mathrm{RI} = \frac{1}{K}\sum_{k=1}^{K}\text{Rep-}n(\mathbf{Y}_k). \tag{3}$$

Information Entropy (IE). Finally, we adopt an entropy-based metric [31] that measures the diversity of token-level n-grams. Given p_i, which denotes the empirical probability of the i-th n-gram in $\mathbf{Y}$, we define the normalized n-gram entropy as:

$$H_{\text{norm}}(\mathbf{Y}) = -\frac{1}{\log_2 N} \sum_{i=1}^{N} p_i \log_2 p_i, \tag{4}$$

where $N = |\mathcal{S}(\mathbf{Y})|$. Lower entropy indicates stronger repetition and less lexical diversity. The dataset-level IE is then computed as:

$$\text{IE} = \frac{1}{K} \sum_{k=1}^{K} H_{\text{norm}}(\mathbf{Y}_k). \tag{5}$$

Each metric captures a distinct aspect of repetition. RR reflects the presence of degenerate looping behavior, RI quantifies the extent of duplication once repetition emerges, and IE captures the loss of lexical diversity in repetitive generation. Together, these metrics allow us to distinguish isolated repetition artifacts from sustained degeneration, which no single metric can capture alone.

3.2 Standardized Video Testbed

We construct a standardized video testbed by randomly sampling 10,000 videos from several widely used public video datasets, including LLaVA-Video-178K [23], NeXT-QA [32], ActivityNetQA [33], and LLaVA-Hound [22]. As illustrated in Figure 2(a), the resulting collection exhibits substantial diversity in both temporal length and semantic content. Video durations range from short clips to moderately long videos of up to 180 seconds, enabling evaluation across different temporal scales, while the content spans a wide variety of everyday scenarios, with categories such as comedy, lifestyle, and sports appearing frequently. This diversity makes the testbed representative of common VideoLLM usage scenarios and well suited for comprehensively evaluating output repetition behavior.

3.3 Temporal Stressor Library

The temporal stressor library is created with a set of temporal transformations that alter the temporal structure of a video while largely preserving its semantic content. Formally, given a video sequence $\mathcal{V} = \{I_1, \ldots, I_T\}$, a temporal transformation $\mathcal{T}$ maps $\mathcal{V}$ to a transformed video sequence:

$$\mathcal{T} : \mathcal{V} \mapsto \mathcal{V}', \quad \mathcal{V}' = \{I'_1, \ldots, I'_{T'}\}, \tag{6}$$

where T' may differ from T depending on the transformation. In VIDEOSTF, we employ five common temporal transformations: Add, Delete, Replace, Reverse, and Shuffle that capture different ways of modifying temporal structure. The five transformations are designed to span a controlled spectrum from *local* perturbations that preserve global temporal coherence to *global* perturbations that destroy it, allowing us to isolate which aspect of temporal structure interacts with output repetition. **Add** randomly selects k frames $\{I_{i_1}, \ldots, I_{i_k}\}$ from $\mathcal{V}$ and inserts them at k random positions $\{q_1, \ldots, q_k\}$, producing $\mathcal{V}' = \{I'_1, \ldots, I'_{T+k}\}$ in which each inserted frame I_{i_j} occupies position q_j and the original frames retain their relative order; this introduces local redundancy. **Delete** removes k randomly chosen frames, yielding $\mathcal{V}' = \{I'_1, \ldots, I'_{T-k}\}$ consisting of the remaining frames in their original relative order, which introduces local omissions. **Replace** randomly selects k frames to be replaced, and for each selected I_{i_m} draws a replacement I_{j_m} uniformly from the remaining frames $\mathcal{V} \setminus \{I_{i_1}, \ldots, I_{i_k}\}$, producing $\mathcal{V}' = \{I'_1, \ldots, I'_T\}$ that preserves the original length but introduces local inconsistencies. Together, Add, Delete, and Replace act locally and leave the global temporal trajectory largely intact. In contrast, **Reverse** inverts the order of all frames to produce $\mathcal{V}' = \{I_T, I_{T-1}, \ldots, I_1\}$, leaving each frame's content unchanged but globally inverting the temporal progression, and **Shuffle** applies a random permutation π so that $I'_t = I_{\pi(t)}$ for $t = 1, \ldots, T$, preserving the set of frames while destroying any coherent ordering. Together, the five transformations cover redundancy, omission, local inconsistency, global inversion, and global disorder, providing a principled basis for stress-testing how each aspect of temporal structure affects output repetition.

Table 1: Repetition rates (%), repetition intensity and information entropy of mainstream VideoLLMs.

Model (LLM)	Frames	RR	RI	IE	Model (LLM)	Frames	RR	RI	IE
LLaVA-Video-7B-Qwen2 (Qwen2)	8	3	0.32	0.87	VideoLLaMA2 (Mistral-7B-Instruct-v0.2)	8	13	0.37	0.84
	16	5	0.33	0.86		16	15	0.38	0.84
	24	10	0.34	0.86		24	16	0.39	0.83
	32	7	0.32	0.86		32	15	0.38	0.83
LLaVA-Video-7B-Qwen2-Video-Only (Qwen2)	8	63	0.47	0.80	ShareGPT4Video (Meta-Llama-3-8B-Instruct)	8	91	0.58	0.75
	16	67	0.48	0.80		16	85	0.57	0.76
	24	61	0.46	0.80		24	79	0.55	0.76
	32	58	0.46	0.81		32	82	0.55	0.77
LLaVA-NeXT-Video-7B (Vicuna-7b-v1.5)	8	15	0.42	0.82	InternVL3.5-8B (Qwen3-8B)	8	9	0.36	0.84
	16	17	0.41	0.82		16	7	0.32	0.86
	24	18	0.42	0.83		24	11	0.37	0.84
	32	-	-	-		32	7	0.35	0.85
LLaVA-NeXT-Video-7B-DPO (Vicuna-7b-v1.5)	8	11	0.39	0.83	Qwen3-VL-8B-Instruct (Qwen3-8B)	8	8	0.36	0.85
	16	8	0.37	0.85		16	13	0.36	0.85
	24	9	0.38	0.84		24	11	0.35	0.85
	32	-	-	-		32	11	0.35	0.84
LLaVA-NeXT-Video-32B-Qwen (Qwen1.5-32B)	8	18	0.35	0.85	Molmo2-8B (Qwen3-8B)	8	65	0.49	0.80
	16	25	0.36	0.85		16	69	0.50	0.79
	24	22	0.36	0.85		24	66	0.49	0.79
	32	29	0.37	0.84		32	72	0.51	0.78

4 Experiments

4.1 Experiment Setup

VIDEOSTF tests 10 VideoLLMs, including LLaVA-Video-7B-Qwen2 [17], LLaVA-Video-7B-Qwen2-Video-Only [17], LLaVA-NeXT-Video-7B [18], LLaVA-NeXT-Video-7B-DPO [18], LLaVA-NeXT-Video-32B-Qwen [18], VideoLLaMA2 [19], ShareGPT4Video [5], InternVL3.5-8B [3], Qwen3-VL-8B-Instruct [20], and Molmo2-8B [21]. All models are tested under deterministic settings, with `do_sample` set to `False` and temperature fixed at 0.0, to avoid decoding randomness.

We adopt the three metrics: Repetition Rate (RR), Repetition Intensity (RI), and Information Entropy (IE) (Section 3.1). For RR, we set $n = 5$, which strikes a balance between sensitivity and robustness: this choice avoids overestimating repetition due to frequent function-word patterns while reliably capturing non-trivial repeated phrases indicative of pathological repetition in VideoLLM outputs. For RI and IE, we set $n = 1$ following [28], as unigram-based statistics are most sensitive to the emergence of repetition, whereas larger n values become increasingly insensitive and less effective in distinguishing repetitive from normal outputs. Nevertheless, VIDEOSTF supports all metrics with arbitrary n, and we report results with other n values in Section B.1.

4.2 Pervasive Testing

Pervasive testing evaluates whether output repetition consistently arises in VideoLLMs with different base LLMs and under varying numbers of sampled frames. We assess 10 VideoLLMs with various base LLMs. For each model, we consider four different numbers of sampled frames: 8, 16, 24, and 32. By default, LLaVA-NeXT-Video-7B and LLaVA-NeXT-Video-7B-DPO do not enforce a fixed number of sampled frames, leading to fewer frames for short videos. To ensure fair comparison, we enable forced frame sampling for both models in the testing.

Table 1 shows that all the 10 VideoLLMs, despite being built on different base LLMs, exhibit output repetition (typical output repetition examples are provided in Section B.2). Among them, ShareGPT4Video and Molmo2-8B display the most severe repetition, with repetition rates exceeding 79% and 65%, respectively. Moreover, across all models, output repetition remains stable across 8, 16, 24, and 32 sampled frames, indicating that this failure mode is insensitive to the commonly used input temporal length. For LLaVA-NeXT-Video-7B and LLaVA-NeXT-Video-7B-DPO, we observe empty outputs when the number of sampled frames reaches 32 because the resulting visual embeddings exceed the maximum input capacity of the underlying LLMs, breaking the generation and preventing repetition from being measured. In summary, these results indicate that output repetition is pervasive in modern VideoLLMs on unperturbed videos, neither tied to a particular underlying LLM nor sensitive to the number of sampled frames within this range.

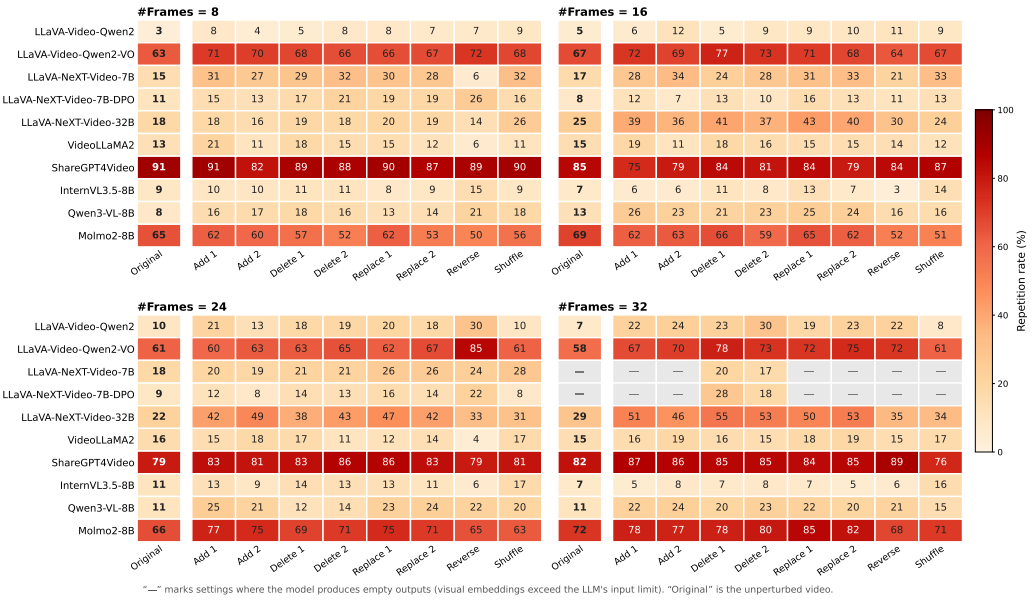

Figure 3: Repetition rates (%) on original videos and under temporal transformations.

4.3 Temporal Stress Testing

In this testing, we consider five temporal transformations, Add Frames, Delete Frames, Replace Frames, Reverse, and Shuffle (Section 3.3). For the former three ones, we evaluate two variants with $k = 1$ and $k = 2$, resulting in six variants: Add 1, Add 2, Delete 1, Delete 2, Replace 1, and Replace 2. When deleting one frame, we exhaustively traverse all sampled frames and remove each frame once. When deleting two frames, we run 30 trials and randomly select two distinct frames to remove in each trial (to balance coverage across different sampled-frame numbers). For Add Frames and Replace Frames, where insertion or replacement positions are stochastic, we perform 30 trials. Finally, for Shuffle, we generate 30 random permutations per video.

Figure 3 shows that temporal transformations increase repetition rates in almost all cases. We summarize this effect by computing, for each sampled frame count, the gap between the mean repetition rate over the eight stressor variants and that of the original videos, normalized by the latter. The relative increase is 133%, 78%, 86%, and 205% at 8, 16, 24, and 32 frames for LLaVA-Video-7B-Qwen2, and 108%, 67%, 83%, and 90% for Qwen3-VL-8B-Instruct, with marginal decreases in only a few cases. As shown in Figure 4, RI and IE exhibit similar trends. Representative examples are provided in Section B.3. These results show that temporal modifications can substantially amplify repetitive generation. Since prevalence is established on unperturbed videos in Section 4.2, this test instead measures sensitivity to temporal change. As in pervasive testing, the number of sampled frames has limited overall impact. A small subset of models (*e.g.,* LLaVA-Video-7B-Qwen2) show increasing repetition as more frames are sampled, suggesting greater susceptibility under denser sequences of visually similar frames.

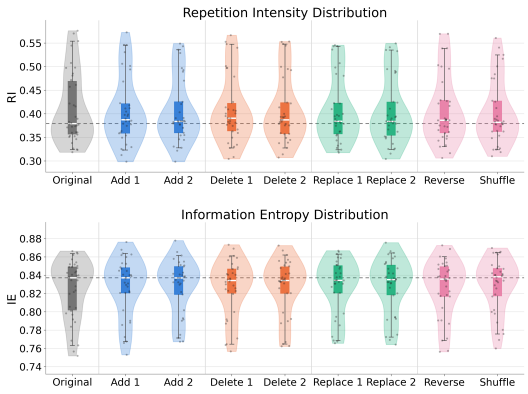

Figure 4: RI and IE distributions under original and temporally transformed inputs across VideoLLMs.

When comparing different temporal transformations, we observe a clear pattern. Add, Delete, and Replace preserve partial temporal coherence while introducing redundancy or local inconsistencies, which can reinforce similar predictions and promote repetition. Reverse instead disrupts temporal order globally, weakening learned temporal priors and reducing repetitive descriptions. Shuffle lies between these extremes: although it destroys global order, its local randomness rarely produces

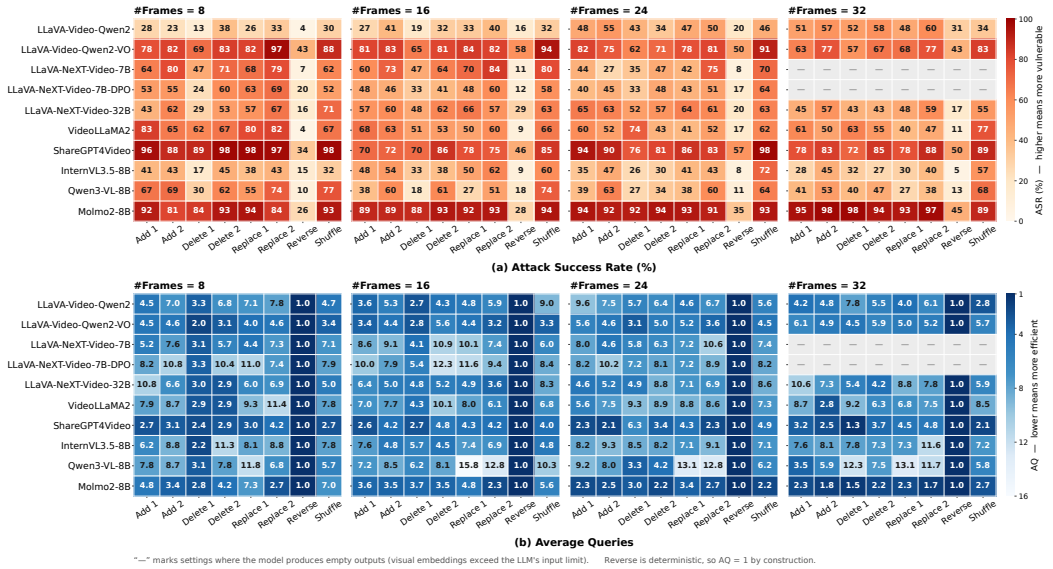

Figure 5: Attack performance of stressors on videos that originally produce non-repetitive outputs.

fully inverted sequences, leading to fragmented but still repetition-prone cues. Because Add and Delete also alter frame count, we compare them with count-matched uniform re-sampling. Add remains stronger than its control, whereas Delete does not, indicating that part of the Delete effect is attributable to frame count rather than temporal disruption. Section C.4 provides more details.

4.4 Adversarial Exploitation

Our temporal stress testing shows that temporal transformations can amplify output repetition in VideoLLMs. We therefore examine whether input-level temporal manipulations can flip videos that originally produce non-repetitive outputs into ones that trigger repetition. We consider a black-box setting, as in prior attacks on video models [34–37] and VideoLLMs [38, 39], in which an attacker can only query the target VideoLLM and apply temporal transformations to the input videos to induce repetition, without access to the model's parameters or internal architecture. For each VideoLLM, we select the video examples from pervasive testing (Section 4.2) that originally produce normal outputs as the attack set. We leverage the temporal transformations defined in Section 3.3 and treat each of them as a distinct attack type. Given a video, we iteratively apply temporal transformations and query the target VideoLLM with each transformed video until output repetition is observed or a preset maximum of 30 query attempts is reached. We evaluate attack effectiveness using two metrics. Attack Success Rate (ASR) measures the proportion of videos whose outputs become repetitive (measured by RR), and Average Queries (AQ) reports the average number of queries required to induce repetition among successfully attacked samples.

Figure 5 shows that for almost all transformations, the attacks succeed with very few queries. The highest ASR reaches 98%, and the maximum AQ remains as low as 15.8. Specifically, Reverse yields the weakest attack performance due to its single and deterministic temporal change. In contrast, the remaining attack types achieve consistently high ASRs. Even on models that exhibit relatively low repetition on original videos, attacks remain highly effective, with VideoLLaMA2 achieving 40% to 83% ASR and InternVL3.5-8B achieving 28% to 72% ASR. These results indicate that output repetition is not a corner case confined to a small subset of videos. Overall, videos that initially produce normal outputs often fail to preserve this behavior under temporal manipulations, revealing a practical vulnerability: with only black-box query access, repetitive degeneration can be induced and, in extreme cases, driven until the output token limit is reached, leading to denial-of-service (DoS).

5 Analysis

Sections 4.2 to 4.4 establish that repetition is pervasive, temporally amplified, and exploitable. We now use VIDEOSTF to examine *why* it emerges and what it costs. Analyses that reuse the settings of the main experiments run on the full testbed, while the additional controls run on an 800-video subset

Table 2: Repetition rates (%) under visual-input controls with fixed weights, prompt, and decoding.

Model	Text only	Black	Noise	Static	Genuine frames						
					1	2	4	8	16	24	32
LLaVA-Video-7B-Qwen2	56	2	0	0	9	2	10	3	7	11	8
LLaVA-Video-7B-Qwen2-Video-Only	85	11	0	30	4	23	50	64	68	64	61
ShareGPT4Video	0	0	0	12	26	57	76	87	85	80	84
Qwen3-VL-8B-Instruct	0	0	0	11	14	8	6	9	12	12	10
Molmo2-8B	0	0	2	27	15	22	21	59	63	60	74

of the testbed, so their absolute rates differ slightly from those of the main experiments. Please refer to Section C for more results and detailed analysis, where Section C.7 further shows that repetition persists on other video distributions.

Severity Spectrum. Severe repetition (token cap reached) is concentrated in the most affected models (*e.g.,* ShareGPT4Video, Molmo2-8B), while ordinary repetition dominates and already degrades quality, and Figure 9 shows that the maximum n-gram count thins rapidly beyond $\tau > 1$. Most repetitions are thus moderate rather than DoS-level, and VIDEOSTF captures the *onset* of repetition.

Repetition and Output Quality. On unperturbed videos in the subset, repetition tracks the information density of a description (distinct content words per 100 words) almost monotonically, with Spearman correlations of -0.95 for RI and 0.92 for IE. Severe outputs are 2.6× longer than benign ones yet contain no more distinct content words, while CLIPScore [40] stays flat when averaged over the sentences of a description. Multiple-choice accuracy on the same NExT-QA [32] clips barely depends on whether the model's description repeats (86.6% *vs.* 83.6% for LLaVA-Video-7B-Qwen2 and 84.7% *vs.* 84.9% for its Video-Only variant). This suggests that repetition is a failure of generation stability rather than of video understanding, which accuracy-based benchmarks [9–11] do not capture, and Section C.2 reports the full results.

Visual-Input Controls. We isolate the effect of visual input by fixing the weights, prompt, and decoding configuration while varying only the video input. As summarized in Table 2, uninformative inputs produce little repetition (0%–11%), while a maximally redundant static video yields less repetition than 16 genuine frames (0%–30% *vs.* 7%–85%). Repetition also increases with frame count only for susceptible models and largely plateaus beyond 8 frames. Although the two LLaVA-Video variants can loop without video by fabricating frame-by-frame narrations, text-only repetition does not predict video repetition across models. Likewise, the shared Qwen3-8B [41] backbone does not determine the behavior of Qwen3-VL-8B-Instruct and Molmo2-8B. Together, these controls show that repetition depends on visual input but is not explained by visual redundancy alone. Instead, it increases with distinct frames to describe, while repeated spans typically re-identify the same people, objects, or scenes. Full analyses are provided in Section C.3.

Local Temporal Disruption. Stressor-induced amplification is not explained by generic input corruption. On LLaVA-Video-7B-Qwen2 at 32 frames in the subset, replacing 2 frames with frames from a different video raises repetition from 8% to 17%, whereas blurring, noising, JPEG-compressing, or masking the same 2 frames changes it by at most 2%, and a 2× speed re-sample that keeps the real order lowers it to 4%, as detailed in Section C.4. Inside the encoder, Figure 6(a) shows that Δrepetition closely tracks Δadjacent-frame similarity across stressors, while global token similarity and effective rank remain essentially unchanged. Figure 6(b) quantifies this, and only adjacent-frame similarity is strongly correlated with Δrepetition (Pearson 0.81 over 8 stressors), whereas global similarity and rank are near zero.

We therefore identify disrupted short-range temporal structure as the feature-level signal most consistently associated with repetition amplification, providing evidence of a strong relationship between local temporal disruption and repetitive behavior. On the decoder side, measured on the subset, the share of attention on visual tokens halves from the first to the last 50 generated tokens, and it does so equally for looping and benign outputs, so a loop does not arise because the decoder stops attending to the video, as shown in Section C.4.

Training Data. An audit of 120,000 captions from six public training corpora shows that the looping vocabulary is learned, since "continues to" appears 41,688 times in 20,000 LLaVA-Video-178K captions, and Section C.5 reports the full audit. Corpus repetitiveness alone does not predict model repetitiveness, as PixMo-Cap [21] has the lowest caption RI yet trains the highly repetitive Molmo2-8B. The training mixture helps reconcile this, because removing non-video data from the mixture

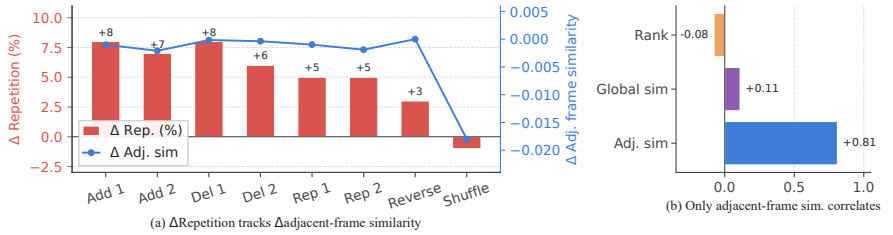

Figure 6: Feature-level changes induced by temporal transformations.

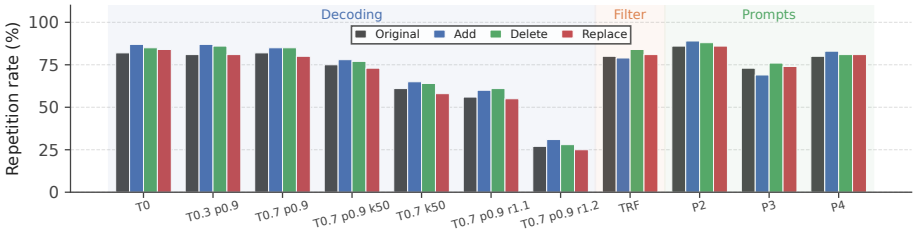

Figure 7: Repetition rates of ShareGPT4Video under decoding strategies, TRF, and prompts.

of LLaVA-Video-7B-Qwen2 moves RI from 0.33 to 0.48, toward the caption RI of 0.677. With the lower repetition of the DPO variant, this indicates that training-stage choices influence the rate, although confirming the link would require retraining with controlled caption length.

Decoding-Time Mitigation. Figure 7 reports ShareGPT4Video at 32 frames on the full testbed and shows that sampling, temporal redundancy filtering (TRF), and rephrased prompts (P2 to P4, described in Section C.6) have limited effect, while temporal stressors amplify repetition under *every* setting. The repetition penalty is the exception. It lowers repetition from 82% to 56% and 27% at penalties of 1.1 and 1.2. To quantify what this costs, we repeat the experiment on LLaVA-Video-7B-Qwen2-Video-Only in the subset and add a stronger penalty of 1.3. Repetition falls from 61% under greedy decoding to 36% and 10% at penalties of 1.2 and 1.3, but CLIPScore drops by 1.7%, 2.7%, and 5.5% at penalties of 1.1, 1.2, and 1.3 while output length does not fall. Multiple-choice and short-answer accuracy stay unchanged across penalty strengths because the answers are too short to repeat, as reported in Section C.6. The penalty is thus an effective symptom-level control whose grounding cost must be tuned, and no setting we test reaches both low repetition and the original grounding. As a test-time remedy that keeps the weights, decoding, and frame budget, re-querying only the outputs flagged by RR with alternative temporal views recovers 41% of repetitive outputs, as described in Section C.6.

6 Conclusion

We present VIDEOSTF, a benchmarking framework for output repetition in VideoLLMs, comprising three n-gram-based metrics, a standardized testbed, and a library of controlled temporal stressors. Together with the analyses it enables, VIDEOSTF identifies output repetition as a pervasive, temporally amplified, and exploitable failure mode that is not explained by visual redundancy, and whose amplification tracks local temporal disruption. Repetition penalties are the only common mitigation that lowers repetition, but the gain comes with weaker visual grounding, so decoding-time control alone is not a satisfactory remedy. Our training-data audit instead points to video-caption supervision as a promising target for mitigation. Limitations and broader impacts are discussed in Section D. We hope VIDEOSTF serves as a foundation on which the community builds toward more robust, temporally stable video-language systems.

Acknowledgments and Disclosures

We thank the reviewers for their constructive comments. This paper was supported in part by National University of Singapore and University of New South Wales. The authors declare no competing financial interests or personal relationships that could have influenced this work.

References

[1] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[2] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[3] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, "Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[4] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 714–10 726.

[5] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, Z. Tang, L. Yuan *et al.*, "Sharegpt4video: Improving video understanding and generation with better captions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 19 472–19 495, 2024.

[6] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[7] X. Tang, J. Qiu, L. Xie, Y. Tian, J. Jiao, and Q. Ye, "Adaptive keyframe sampling for long video understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 118–29 128.

[8] K. Gao, Y. Bai, J. Gu, S.-T. Xia, P. Torr, Z. Li, and W. Liu, "Inducing high energy-latency of large vision-language models with verbose images," in *Proceedings of the International Conference on Learning Representations*, 2024.

[9] M. Ning, B. Zhu, Y. Xie, B. Lin, J. Cui, L. Yuan, D. Chen, and L. Yuan, "Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models," *Computational Visual Media*, 2025.

[10] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 195–22 206.

[11] Z. Li, X. Wu, H. Du, H. Nghiem, and G. Shi, "Benchmark evaluations, applications, and challenges of large vision language models: A survey," *arXiv preprint arXiv:2501.02189*, vol. 1, 2025.

[12] H. Wu, D. Li, B. Chen, and J. Li, "Longvideobench: A benchmark for long-context interleaved video-language understanding," *Advances in Neural Information Processing Systems*, vol. 37, pp. 28 828–28 857, 2024.

[13] M. Cao, P. Hu, Y. Wang, J. Gu, H. Tang, H. Zhao, C. Wang, J. Dong, W. Yu, G. Zhang *et al.*, "Video simpleqa: Towards factuality evaluation in large video language models," *arXiv preprint arXiv:2503.18923*, 2025.

[14] Y. Cao, W. Song, D. Wang, J. Xue, and J. S. Dong, "Failures to surface harmful contents in video large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 42, 2026, pp. 35 331–35 339.

[15] C. Li, E. W. Im, and P. Fazli, "Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 723–13 733.

[16] W. Zhou, M. Hendy, S. Yang, Q. Yang, Z. Guo, Y. Luo, L. Hu, and D. Wang, "Flattery in motion: Benchmarking and analyzing sycophancy in video-llms," *arXiv preprint arXiv:2506.07180*, 2025.

[17] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[18] Y. Zhang, B. Li, h. Liu, Y. j. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li, "Llava-next: A strong zero-shot video understanding model," April 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/

[19] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao *et al.*, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024.

[20] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[21] C. Clark, J. Zhang, Z. Ma, J. S. Park, M. Salehi, R. Tripathi, S. Lee, Z. Ren, C. D. Kim, Y. Yang *et al.*, "Molmo2: Open weights and data for vision-language models with video understanding and grounding," *arXiv preprint arXiv:2601.10611*, 2026.

[22] R. Zhang, L. Gui, Z. Sun, Y. Feng, K. Xu, Y. Zhang, D. Fu, C. Li, A. Hauptmann, Y. Bisk *et al.*, "Direct preference optimization of video large multimodal models from language model reward," *arXiv preprint arXiv:2404.01258*, 2024.

[23] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 5971–5984.

[24] Y. Li, C. Wang, and J. Jia, "Llama-vid: An image is worth 2 tokens in large language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 323–340.

[25] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li *et al.*, "Video-ollama 3: Frontier multimodal foundation models for image and video understanding," *arXiv preprint arXiv:2501.13106*, 2025.

[26] Z. Fu, W. Lam, A. M.-C. So, and B. Shi, "A theoretical analysis of the repetition problem in text generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 12 848–12 856.

[27] H. Li, T. Lan, Z. Fu, D. Cai, L. Liu, N. Collier, T. Watanabe, and Y. Su, "Repetition in repetition out: Towards understanding neural text degeneration from the data perspective," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72 888–72 903, 2023.

[28] J. Yao, S. Yang, J. Xu, L. Hu, M. Li, and D. Wang, "Understanding the repeat curse in large language models from a feature perspective," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025.

[29] W. B. Cavnar, J. M. Trenkle *et al.*, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175. Las Vegas, NV, 1994, p. 14.

[30] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," in *Proceedings of the International Conference on Learning Representations*, 2020.

[31] D.-Y. Tsai, Y. Lee, and E. Matsuyama, "Information entropy measure for evaluation of image quality," *Journal of digital imaging*, vol. 21, no. 3, pp. 338–347, 2008.

[32] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9777–9786.

[33] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9127–9134.

[34] L. Jiang, X. Ma, S. Chen, J. Bailey, and Y.-G. Jiang, "Black-box adversarial attacks on video recognition models," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 864–872.

[35] S. Xie, H. Wang, Y. Kong, and Y. Hong, "Universal 3-dimensional perturbations for black-box attacks on video recognition systems," in *IEEE Symposium on Security and Privacy*. IEEE, 2022, pp. 1390–1407.

[36] Y. Cao, X. Xiao, R. Sun, D. Wang, M. Xue, and S. Wen, "Stylefool: Fooling video classification systems via style transfer," in *IEEE symposium on security and privacy*. IEEE, 2023, pp. 1631–1648.

[37] D. Tang, Y. Cao, X. Xiao, D. Wang, S. Wen, and T. Zhu, "Query-efficient video adversarial attack with stylized logo on service computing," *IEEE Transactions on Dependable and Secure Computing*, 2026.

[38] Y. Cao, W. Song, J. Xue, and J. S. Dong, "Poisoning prompt-guided sampling in video large language models," *arXiv preprint arXiv:2509.20851*, 2025.

[39] H. Liu, K. Gao, Y. Bai, J. Li, J. Shan, T. Dai, and S.-T. Xia, "Protecting your video content: Disrupting automated video-based llm annotations," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24 056–24 065.

[40] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7514–7528.

[41] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[42] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[43] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[44] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "Sharegpt4v: Improving large multi-modal models with better captions," in *European Conference on Computer Vision*. Springer, 2024, pp. 370–387.

[45] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[46] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2020.

[47] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 889–898.

[48] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," *arXiv preprint arXiv:1909.05858*, 2019.

[49] W. Song, Y. Sui, Z. Xing, L. Zhu, and J. Xue, "Vidtoken: A video-transformer-based latent token defense for adversarial video detection," in *International Symposium on Research in Attacks, Intrusions and Defenses*. IEEE, 2025, pp. 566–582.

[50] Y. Chen, T. Cao, H. Li, Y. Liu, Y. Li, Y. He, L. M. Khoi, Y. Song, S. Yan, and B. Hooi, "Webagentguard: A reasoning-driven guard model for detecting prompt injection attacks in web agents," *arXiv preprint arXiv:2604.12284*, 2026.

[51] Y. Chen, H. Li, Y. Sui, Y. Liu, Y. He, X. Bai, C. Fei, L. Yabo, H. Ma, Y. Song, and B. Hooi, "Robustness via referencing: Defending against prompt injection attacks by referencing the executed instruction," in *Findings of the Association for Computational Linguistics: ACL 2026*, 2026, pp. 1194–1215.

[52] W. Song, C. Cong, H. Zhong, and J. Xue, "Correction-based defense against adversarial video attacks via discretization-enhanced video compressive sensing," in *USENIX Security Symposium*, 2024, pp. 3603–3620.

[53] A. Robey, E. Wong, H. Hassani, and G. J. Pappas, "Smoothllm: Defending large language models against jailbreaking attacks," *arXiv preprint arXiv:2310.03684*, 2023.

[54] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025, pp. 24 108–24 118.

[55] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 995–19 012.

[56] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," *arXiv preprint arXiv:1910.01442*, 2019.

[57] Q. Li, R. Yu, and X. Wang, "Vid-sme: Membership inference attacks against large video understanding models," *arXiv preprint arXiv:2506.03179*, 2025.

Appendix

A Computational Resource

All experiments are conducted on two RTX 4090 GPUs, which are sufficient to support the evaluated models. As our benchmark is designed to be extensible to future and more complex VideoLLMs, reproducibility should follow the computational requirements specified by each individual model.

B More Results on Main Experiments

This section presents additional results for the main experiments, covering the sensitivity of VIDEOSTF to different choices of n, representative repetition examples across VideoLLMs, and qualitative results under temporal stressors. To improve the layout and overall presentation of the appendix, the figures associated with Section B.2 and Section B.3 are placed together at the end of the appendix, following all other sections.

B.1 Repetition Results with Different n

Different choices of n exhibit varying sensitivity to repetition. We evaluate all three metrics, RR, RI, and IE (Section 3.1) with n ranging from 1 to 10. As shown in Figure 8, smaller n values (*e.g.,* $n \leq 2$) yield higher RR but tend to overcount common short phrases, whereas larger n values increasingly fail to capture repetitive patterns, especially when $n \geq 7$. For RI and IE, repetition becomes difficult to observe when $n \geq 4$. Importantly, the relative repetition tendencies across models remain stable across different n values, with ShareGPT4Video, LLaVA-Video-7B-Qwen2-Video-Only and Molmo2-8B exhibiting higher repetition. In general, although we use $n = 5$ for RR and $n = 1$ for RI and IE in the main paper, chosen to balance sensitivity and robustness for RR and to provide higher sensitivity to repetition for RI and IE, the same repetition trends and cross-model comparisons remain consistent across other choices of n.

B.2 Examples of Output Repetition on Different VideoLLMs

Figures 12, 13, 14, 15, and 16 present representative examples of repetitive outputs from the 10 VideoLLMs evaluated in this paper. As reported in the main paper, repetition is observed across all the 10 VideoLLMs, suggesting that it is a pervasive phenomenon. Meanwhile, the extent of repetition varies across models, resulting in different repetition behaviors. Models such as LLaVA-Video-7B-Qwen2-Video-Only and ShareGPT4Video often exhibit consecutive sentence-level duplication, whereas models such as LLaVA-Video-7B-Qwen2 and VideoLLaMA2 more frequently repeat shorter phrases. In addition, since LLaVA-NeXT-Video-7B and LLaVA-NeXT-Video-7B-DPO produce empty outputs at 32 frames, we further examine whether increasing the number of sampled frames correlates with abnormal generation. We find that when the sampled frame number falls within $[28, 32)$, some videos still yield empty outputs, while many others exhibit highly repetitive responses that often contain grammatical errors, as shown in Figure 13(c). These observations indicate that these models exhibit limited decoding stability and robustness under denser frame sampling settings.

B.3 Examples under Temporal Stressors

Figures 17, 18, 19, and 20 present a representative video example, showing sampled frames and corresponding model outputs for the original video and its transformed counterparts produced by the temporal transformations in the temporal stressor library of VIDEOSTF. The source video depicts a train arriving at a station. Substantial repetition is observed in the generated outputs not only for the original video but also across all transformed videos. Moreover, under several transformations, such as Add 1 and Shuffle, the repetition becomes more severe than that of the original video. This example provides supporting evidence that temporal transformations can amplify repetitive generation in VideoLLMs even when the semantic content of the video is preserved.

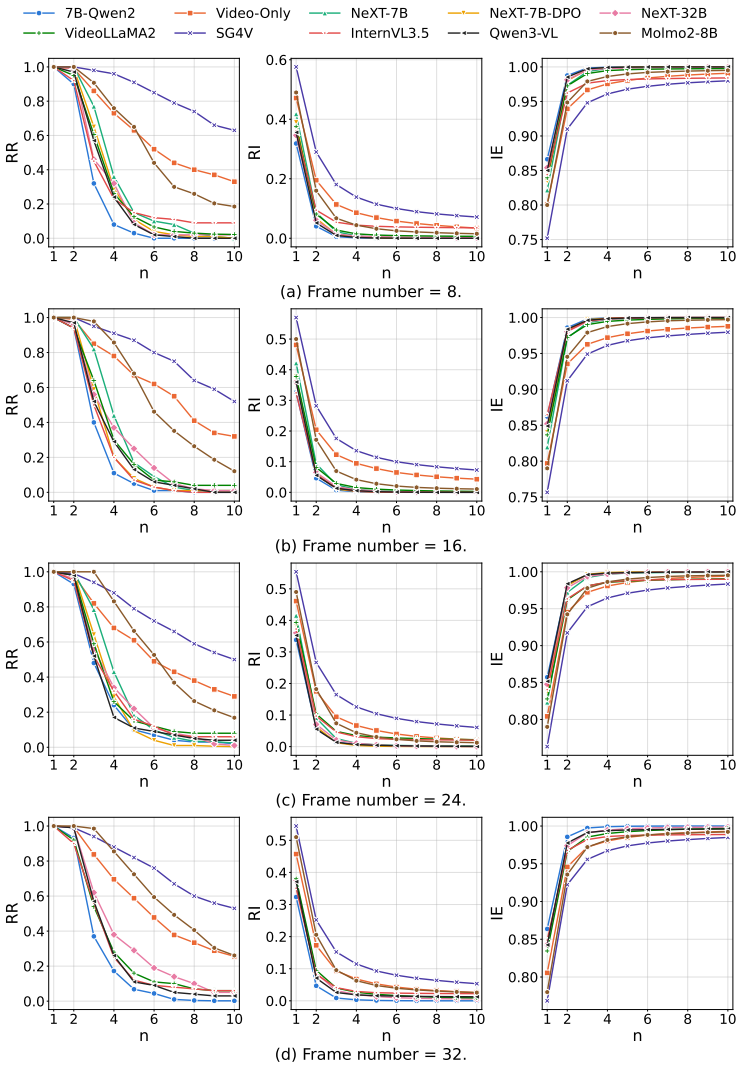

Figure 8: Repetition results of different VideoLLMs under three metrics across varying n.

C More Details and Results on Analysis

This section supplements Section 5 with the experimental details, full results, and extended discussion behind each of its analyses. As in Section 5, the analyses that follow the settings of the main experiments are conducted on the full testbed, whereas the additional controls are conducted on an 800-video subset of it, so their absolute rates deviate slightly from those in Section 4.

C.1 Severity Spectrum

Repetition naturally exhibits different levels, including ordinary repetition and severe repetition (token cap reached). Ordinary repetition degrades user experience and response quality, while severe repetition leads to pathological degenerate loops that can exhaust the decoding budget. We report proportions and average token volume of benign outputs, ordinary repetition, and severe repetition in Table 3. Ordinary repetition has moderate token length, while severe cases consistently reach the decoding limit (1024), showing repetition spans a spectrum, and only a subset corresponds to DoS-like behavior. Output length also matters: shorter outputs rarely become severe, but once a degenerate loop starts, length keeps growing, even when the limit is extended to 4096.

Our repetition metric ($n = 5$, $\tau = 1$) is designed to capture the *onset of repetition*, rather than only extreme cases. We therefore conduct ablations with stricter thresholds (increasing the required

repetition count). Even under higher thresholds, a substantial fraction of outputs remain repetitive, indicating that our conclusions do not depend on a weak definition or incidental phrasing.

Table 3: Repetition severity statistics (%) across models at 32 frames. Rep denotes repetition, Avg Tok is the average number of output tokens, and column $> k$ gives the share of outputs whose maximum n-gram count exceeds k.

Model	Benign	Avg Tok	Ordinary Rep	Avg Tok	Severe Rep	Avg Tok	>1	>2	>3	>4
LLaVA-Video-7B-Qwen2	93	131.54	7	188.29	0	-	7	1	0	0
LLaVA-Video-7B-Qwen2-Video-Only	42	171.5	51	253.83	7	1024	58	27	14	8
ShareGPT4Video	18	206.06	65	356.82	17	1024	82	47	22	11
VideoLLaMA2	85	124.87	13	173.23	2	1024	15	7	3	3
Molmo2-8B	28	326.58	49	403.94	23	1024	72	32	13	9

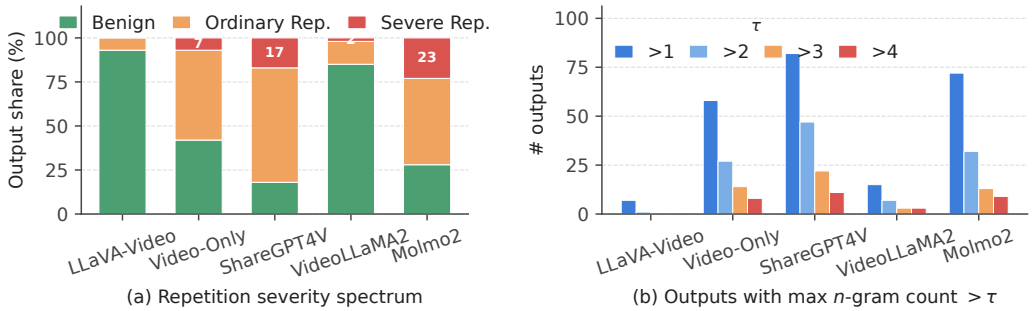

Figure 9: Repetition severity statistics across models. (a) Distribution of benign, ordinary, and severe (token-cap) outputs. (b) Number of outputs whose maximum n-gram count exceeds τ.

C.2 Repetition and Output Quality

On the subset, we correlate quality measures with all three VIDEOSTF metrics on unperturbed videos across LLaVA-Video-7B-Qwen2, LLaVA-Video-7B-Qwen2-Video-Only, ShareGPT4Video, Qwen3-VL-8B-Instruct, and Molmo2-8B. Because the testbed is built on NExT-QA [32], which provides question-answer pairs rather than reference captions, we adopt reference-free quality measures: the number of distinct content words, information density (distinct content words per 100 words), and CLIPScore [40]. Table 4 reports their correlations.

Table 4: Spearman correlation (ρ) between quality measures and the repetition metrics.

Quality measure	RR	RI	IE
Information density	-0.76	-0.95	0.92
Output length (words)	0.72	0.84	-0.84
Distinct content words	0.35	0.40	-0.41
CLIPScore	0.00	0.03	-0.04

Repetition is strongly associated with information density, with RI at -0.95 and IE at 0.92. RR is weaker at -0.76 because it is a thresholded rate rather than a graded quantity, which motivates reporting all three VIDEOSTF metrics. Output length increases sharply with repetition while the number of distinct content words changes little, indicating that longer captions carry limited additional information. CLIPScore remains nearly unchanged. This is expected because it averages over the sentences in a caption, so repeating a sentence that already matches the video has little effect on the average. Table 5 shows the same pattern when outputs are grouped by the severity categories in Table 3. Severe outputs are 2.6× longer than benign ones yet contain no more distinct content words (64.7 *vs.* 67.0), while information density drops by 63%, from 46.1 to 16.9 distinct content words per 100 words (Mann-Whitney U test [42], $p = 1.1 \times 10^{-155}$). Thus, severe repetition uses roughly 2.6× the tokens without adding any new information to the description.

Answer accuracy behaves differently. NExT-QA provides five-way multiple-choice questions on the same clips, covering causal, temporal, and descriptive aspects. At 32 frames on this subset, we split the accuracy of each model by whether its free-form description of the clip is repetitive, as reported in Table 6. We use two models spanning Table 1, namely LLaVA-Video-7B-Qwen2 and its Video-Only variant, whose RR on the full testbed at 16 frames is 5% and 67%. For both models, accuracy is

Table 5: Output quality by repetition severity.

Group (maximum n-gram count)	Share of outputs (%)	Words	CLIPScore	Distinct content words	Information density
Benign (1)	72.2	145	0.2163	67.0	46.1
Ordinary (2 to 4)	24.7	230	0.2232	78.8	34.3
Severe (more than 4)	3.1	382	0.2198	64.7	16.9

nearly unchanged across the two groups. LLaVA-Video-7B-Qwen2 is 3.0% more accurate on videos it describes without repetition, while the Video-Only variant is 0.2% more accurate on videos it describes repetitively, so the direction of the difference is inconsistent. A model can therefore enter a loop during free-form description while retaining its usual multiple-choice accuracy on the same video. This suggests that repetition reflects a failure of generation stability rather than a general failure of video understanding. Multiple-choice accuracy cannot capture this behavior because the output is restricted to a single option and cannot loop. Together, the quality correlations and accuracy split show that the VIDEOSTF metrics capture the stability and informativeness of descriptions, an axis not covered by accuracy-based benchmarks [9–11].

Table 6: NExT-QA multiple-choice accuracy (%) split by whether the description repeats.

Model	Benign description	Repetitive description
LLaVA-Video-7B-Qwen2	86.6	83.6
LLaVA-Video-7B-Qwen2-Video-Only	84.7	84.9

C.3 Visual-Input Controls

The metrics of VIDEOSTF quantify repetition in the output text, so whether the observed repetition is induced by video input or inherited from the base LLM must be established experimentally. In every control below, the decoder weights, prompt, and decoding configuration are fixed, while only the visual input is varied. Any resulting change in RR can therefore be attributed to the visual input. All controls are conducted on the subset, and Section 5 defines each setting and summarizes the results in Table 2, which we analyze in more detail below.

Uninformative visual input produces little repetition. Black frames yield 0% to 11% RR and noise frames yield 0% to 2%, despite supplying visual tokens without meaningful temporal content. A static video with maximal redundancy yields 0% to 30%, compared with 7% to 85% for 16 genuine frames, and no model repeats more on the static input than on genuine frames. The failure therefore requires multiple distinct real frames. Increasing frame count amplifies repetition only in susceptible models. For the three high-repetition models, LLaVA-Video-7B-Qwen2-Video-Only, ShareGPT4Video, and Molmo2-8B, RR rises from 4%, 26%, and 15% at one frame to 61%, 84%, and 74% at 32 frames, with most of the increase occurring by 8 frames. In contrast, the two low-repetition models, LLaVA-Video-7B-Qwen2 and Qwen3-VL-8B-Instruct, remain between 2% and 14% across all frame counts. Table 1 covers 8, 16, 24, and 32 frames, where these curves have largely plateaued, so the frame-count stability reported in Section 4.2 remains valid for these commonly used settings.

The two LLaVA-Video variants also degenerate when asked to describe a video without visual input, reaching 56% and 85% RR. Inspection shows that they confabulate a frame-by-frame narration before entering a loop. This ungrounded-generation failure is distinct from the behavior measured by VIDEOSTF. Supplying the video reduces RR for LLaVA-Video-7B-Qwen2 from 56% to 7% at 16 frames. By contrast, ShareGPT4Video, Qwen3-VL-8B-Instruct, and Molmo2-8B all have 0% RR without video, yet two become among the most repetitive models with video input. The ranking of the five models by text-only RR is therefore largely unrelated to their ranking under video input. We also evaluate the base LLM itself because the text-only setting still uses the VideoLLM with the visual input removed. Qwen3-8B [41], the backbone of both Qwen3-VL-8B-Instruct and Molmo2-8B, repeats on 46% of the same prompts as a pure text LLM. Yet both derived models have 0% RR without video and reach 12% and 63% under 16 video frames, showing that a shared backbone with a strong repetition tendency does not determine the behavior of the corresponding video models.

We further reduce visual information by retaining a random subset of visual tokens in their original order before decoding for LLaVA-Video-7B-Qwen2-Video-Only at 32 frames. RR decreases from 61% with all tokens to 53% with half and 45% with one quarter, consistent with the frame-count

result that repetition weakens as the decoder receives less visual content. Arranging the same 16 frames into a 4×4 image grid yields 54% RR, compared with 68% when the frames are provided as a video, indicating that the image format reduces but does not eliminate repetition. The amount of visual content therefore accounts for much of the effect, while the video representation further amplifies it. We do not include image-language models as baselines because VIDEOSTF targets VideoLLMs, and repetition in image-language models is outside the scope of this study.

Input Redundancy. A natural hypothesis is that repetition is provoked by visually redundant inputs. As shown in Figure 10, CLIP-based frame similarity is comparable between videos that produce benign and repetitive outputs. Because CLIP frame similarity is computed over frame pairs, it mainly captures appearance-level redundancy and is insensitive to frame order. We therefore complement it with the order-sensitive encoder-level statistics in Section C.4 and the static-video control above. In that control, repeating a single frame 16 times consistently produces less repetition than using 16 genuine frames across all models, showing that maximal visual redundancy alone does not induce stronger repetition. Together with the stressor and attack results, these findings indicate that repetition remains input-dependent but is not explained simply by visual redundancy.

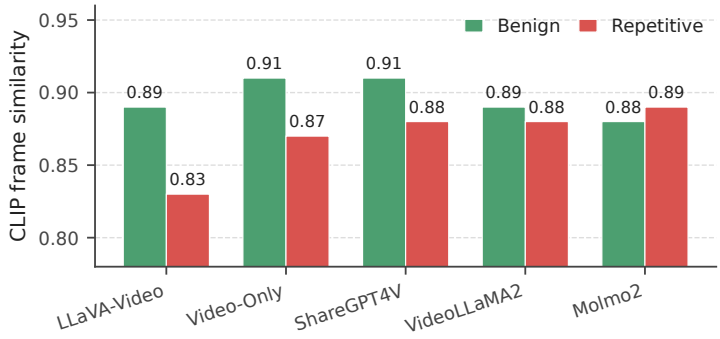

Figure 10: CLIP-based frame similarity of videos with benign *vs.* repetitive outputs.

Repetition on Unperturbed Videos. The mechanism analysis in Section 5 focuses on stressor-induced amplification, so we separately examine why repetition occurs without any stressor and keep the two settings distinct throughout. Prevalence is measured on unperturbed videos alone, where Table 1 reports RR reaching 67%, 91%, and 72% for the Video-Only variant, ShareGPT4Video, and Molmo2-8B, while the temporal stress test measures sensitivity and exploitability.

On the text side, the repeated span is almost always a re-identification phrase. Most often, it names a person or object again using a color or clothing attribute, as in "the person on the right", "woman in the green shirt", or "in a white shirt and". A smaller but substantial group instead repeats references to a scene or spatial arrangement. Across the outputs of LLaVA-Video-7B-Qwen2, LLaVA-Video-7B-Qwen2-Video-Only, ShareGPT4Video, Qwen3-VL-8B-Instruct, and Molmo2-8B on the subset, the most frequent content words within repeated spans are white (2.9% of content-word occurrences), shirt (2.7%), red (2.6%), black (2.3%), person (2.2%), and toy (1.6%). No specific entity or scene consistently triggers repetition. Instead, loops form around repeated re-identification of the same entities and scenes across frames. The static-video control isolates this factor because 16 identical frames use the same frame budget and number of visual tokens as 16 genuine frames but introduce nothing new to re-identify. This reduces RR from 68% to 30% for the Video-Only variant. Repetition on clean videos is therefore associated with how often the description re-identifies the same content, consistent with the frame-count dependence in Table 2.

A clip of two puppies illustrates this failure. LLaVA-Video-7B-Qwen2-Video-Only begins fluently but then cycles between the same events, with the 5-gram "in the corner, facing each other" appearing five times within a single description.

> *The puppies sit in the corner, facing each other, with the bowl now empty. One puppy starts to move towards the other, and they engage in playful behavior, nuzzling and rolling around. The video continues with the puppies sitting in the corner, facing each other, and then moving towards the bowl. One puppy drinks from the bowl while the other watches. They then sit back in the corner, facing each other, and continue their playful interaction. The video*

concludes with the puppies sitting in the corner, facing each other, and then moving towards the bowl. One puppy starts to drink from the bowl while the other watches. They then sit back in the corner, facing each other, and continue their playful interaction.

The output remains grammatical and relevant but stops adding information, repeatedly cycling between two previously described events while re-identifying the same subjects and scene. On the video side, we split the subset at the median of four per-video attributes, namely frame self-similarity, motion magnitude, shot-change count, and duration, and compare RR between the two halves. No attribute shows a consistent direction across models. Shot-change count raises RR by 24% on ShareGPT4Video but lowers it by 16% on Molmo2-8B, while frame self-similarity, the attribute closest to visual redundancy, yields +16% on ShareGPT4Video and -4% on Qwen3-VL-8B-Instruct. Neither the loop content nor these video attributes therefore identifies content that consistently induces repetition across the evaluated models.

C.4 Local Temporal Disruption

Temporal Disruption Versus Generic Corruption. The stressors may increase repetition simply by creating unnatural inputs. To distinguish temporal disruption from generic corruption, we evaluate LLaVA-Video-7B-Qwen2 at 32 frames on the subset. We replace 2 of 32 frames with frames from a different video, or corrupt the same 2 frames using Gaussian blur, Gaussian noise, JPEG compression at quality 10, or cutout [43], so that only the type of perturbation changes. Table 7 reports the results. We also include a $2\times$ speed re-sample that preserves the original frame order and introduces no temporal jump between consecutive frames.

Table 7: Repetition rates (%) of LLaVA-Video-7B-Qwen2 at 32 frames under frame-level changes.

Condition	RR	Change
Original (unperturbed)	8	–
2 frames replaced with frames from a different video	17	+9
Gaussian blur on the same 2 frames	10	+2
Gaussian noise on the same 2 frames	9	+1
JPEG compression (quality 10) on the same 2 frames	8	0
Cutout on the same 2 frames	8	0
$2\times$ speed re-sample (real order, no jump)	4	-4

Blur, noise, JPEG compression, and cutout make the selected frames less natural but have little effect on RR. Only replacement with content from another video clearly increases repetition despite modifying the same number of frames. This pattern points to semantic discontinuity rather than generic corruption, while the lower RR under the order-preserving re-sample supports the same interpretation. Repetition on clean videos also argues against unnatural inputs alone, since Table 2 shows that repetition reaches its full level without any stressor as more frames are added.

Add and Delete modify both the frame count and the temporal structure, while Table 2 shows that frame count itself affects repetition. We therefore compare each stressor with a uniform re-sample at the same frame count on the subset, which changes the number of frames without introducing discontinuity. Table 8 reports the results. Add exceeds its count-matched control by 8% and 6% on the two models, indicating an additional effect from temporal disruption beyond the two extra frames. Delete is 4% below its matched control on both models, indicating that its effect is explained by the sampling change. Cross-video replacement, which preserves the frame count, provides the clearest separation from the unperturbed condition.

Table 8: Repetition rates (%) of Add and Delete against count-matched uniform re-sampling.

Condition	LLaVA-Video-7B-Qwen2	LLaVA-Video-7B-Qwen2-Video-Only
Original, 32 frames	8	61
Add 2, 34 frames	18	55
Uniform re-sample, 34 frames	10	49
Delete 2, 30 frames	10	57
Uniform re-sample, 30 frames	14	61

Encoder Token Statistics. To examine whether temporal redundancy introduced by transformations leads to repetitive outputs, we conduct additional experiments that analyze redundancy at the encoder

token level after temporal transformations, rather than at the frame similarity level. Specifically, we measure three quantities: (1) adjacent-frame token similarity, capturing local temporal redundancy; (2) global token similarity, reflecting global temporal redundancy; and (3) effective rank, measuring token diversity. We then evaluate their changes under each transformation and assess their correlation with the corresponding increase in repetition rate.

Our results show that: (1) although repetition increases under most temporal perturbations, the redundancy signals remain extremely small in magnitude; (2) global token similarity and effective rank show very weak correlation with repetition increase (Pearson corr: 0.11 and -0.076), indicating that increased global redundancy is not the driving factor; (3) adjacent-frame token similarity slightly decreases in absolute terms, but its relative change is strongly correlated with repetition (Pearson corr: 0.81). These results suggest that stressors raise repetition by reshaping short-range temporal token interactions, which may destabilize cross-frame grounding, rather than by making inputs more redundant. Because the analysis correlates changes over eight stressors, it identifies the feature-level signal that tracks amplification and does not by itself establish a cause.

It is also worth noting that our temporal transformations, Add, Delete, and Replace, modify only one or two frames per video and do not create highly repetitive inputs. The insertion positions are random, making it very unlikely to produce consecutive identical frames. This also explains why adjacent-frame token similarity slightly decreases in absolute terms, as temporal perturbations disrupt learned motion continuity and introduce feature-level misalignment between neighboring frames, rather than making adjacent frames more similar. For Reverse and Shuffle, although they globally reorder frames, they do not introduce duplicate or repetitive content and instead break temporal coherence more drastically, so a decrease in adjacent-frame token similarity is also expected.

Table 9: Feature-level changes and their correlation with repetition.

Trans	Δ Rep	Δ Adj Sim	Δ Glob Sim	Δ Rank
Add 1	0.08	-0.0010	8.1E-05	-0.0036
Add 2	0.07	-0.0021	-3.7E-05	-0.00070
Delete 1	0.08	-0.00016	1.9E-05	-0.0038
Delete 2	0.06	-0.00037	6.1E-05	-0.0092
Replace 1	0.05	-0.00099	-0.00017	-0.0065
Replace 2	0.05	-0.0019	0.00025	-0.045
Reverse	0.03	-1.8E-08	6.0E-09	4.3E-07
Shuffle	-0.01	-0.018	3.1E-09	-6.6E-07
Corr with Δ Rep	1.00	0.81	0.11	-0.076

Decoder Attention. As the decoder-side counterpart to Figure 6, we record, for each generated token, the share of attention mass assigned to visual tokens, averaged across all heads and layers. We compare repetitive and benign outputs using LLaVA-Video-7B-Qwen2-Video-Only at 32 frames on the subset. As shown in Table 10, visual attention drops by roughly half from the first to the last 50 generated tokens, with nearly identical trends for repetitive and benign outputs. Measured from the onset of repetition, attention decreases from 15.4% to 10.0%, a smaller decline than that observed for benign outputs over the same number of tokens. The decoder therefore continues to attend to the video during repetition, indicating that the failure is not explained by a loss of visual attention. Instead, the signal most consistently associated with amplification remains the disruption of short-range temporal structure between adjacent frames on the encoder side.

Table 10: Share of decoder attention on visual tokens (%).

Output	First 50 tokens	Last 50 tokens
Repetitive	19.3	9.6
Benign	19.0	9.5

C.5 Training Data

Training Data Audit. To examine the role of training data, we audit 120,000 captions, with 20,000 sampled from each of six public training corpora, namely LLaVA-Video-178K [17], the ShareGPT4Video mix [5], ShareGPT4V [44], LLaVA-Hound [22], LLaVA-Instruct-150K [45], and

PixMo-Cap [21], and apply the VIDEOSTF metrics directly to the caption text. Table 11 reports the results. We report RI rather than RR for the corpora because their mean caption lengths differ by up to 7×, which confounds cross-corpus comparisons of RR. The phrase rate counts occurrences of "continues to" per 1,000 words. These corpus-level statistics are computed on the training captions rather than the testbed. For reference, the table also lists the corresponding models and their full-testbed results at 16 frames from Table 1.

Table 11: Caption repetition in training corpora and the corresponding models.

Training corpus	Mean words	Caption RI	Phrase rate	Model	RR (%)	RI
LLaVA-Video-178K	1055	0.677	1.98	LLaVA-Video-7B-Qwen2	5	0.33
				LLaVA-Video-7B-Qwen2-Video-Only	67	0.48
ShareGPT4Video mix	288	0.387	0.47	ShareGPT4Video	85	0.57
ShareGPT4V	144	0.370	0.02	VideoLLaMA2	15	0.38
LLaVA-Hound	149	0.322	0.00	LLaVA-NeXT-Video-7B-DPO	8	0.37
LLaVA-Instruct-150K	170	0.317	0.03	VideoLLaMA2	15	0.38
PixMo-Cap	168	0.300	0.01	Molmo2-8B	69	0.50

The looping vocabulary is also reflected in the training data. Phrases that recur in model outputs are pervasive in the training targets. Among 20,000 LLaVA-Video-178K captions, "continues to" appears 41,688 times, or roughly twice per caption, "the video continues" appears 22,299 times, and "the scene shifts" appears 14,823 times. Per 1,000 words, this corpus is about 4× denser in these phrases than any other corpus we examine, showing that models reproduce training patterns at the phrase level. Corpus repetitiveness alone, however, does not predict model repetitiveness. Caption RI ranks the corpora as LLaVA-Video-178K (0.677), ShareGPT4Video mix (0.387), ShareGPT4V (0.370), LLaVA-Hound (0.322), LLaVA-Instruct-150K (0.317), and PixMo-Cap (0.300), but the corresponding models do not follow this ordering. The most repetitive corpus trains the least repetitive model in Table 1, while PixMo-Cap has the lowest RI and almost no looping phrases yet trains Molmo2-8B, which reaches RR 69% at 16 frames.

The training mixture helps explain this discrepancy. LLaVA-Video-7B-Qwen2 and its Video-Only variant share the same backbone and video corpus and differ only in whether non-video data is included in the training mixture. Their RI is 0.33 and 0.48 against a caption RI of 0.677, so removing the additional data shifts the model markedly toward the repetitive style of its video captions. Together with the DPO comparison in Table 1, where RR is 15% to 18% without DPO and 8% to 11% with it for otherwise identical models, this pattern indicates that training-stage choices influence repetition more strongly than the base LLM alone. Cleaning the caption corpus by itself is not sufficient, as the PixMo-Cap case shows, so training-level intervention remains a promising direction rather than a complete solution to output repetition.

Caption length provides an additional connection. RR flags an output whenever any 5-gram repeats, so longer text is more likely to be flagged. To control for this effect, we retain only texts that reach W words, truncate them to the first W words, and report the share that still contains a repeated 5-gram in Table 12. The model outputs are taken from the subset. Shuffling the words within each truncated text gives a chance baseline of 0.0% to 0.1% across all settings.

Table 12: Length-matched repetition rates (%) of training captions and model outputs.

Source	Name	Mean words	RR at 150 words	RR at 300 words
Corpus	LLaVA-Video-178K	1055	25.7	86.4
	ShareGPT4Video mix	288	2.7	12.4
	ShareGPT4V	144	11.2	–
Model	LLaVA-Video-7B-Qwen2	–	30.0	–
	LLaVA-Video-7B-Qwen2-Video-Only	–	59.4	100.0

Longer captions exhibit more repetition within every corpus, and even after matching length, repetition rates differ by up to 10× across corpora. At 150 words, both models trained on LLaVA-Video-178K repeat more often than the corpus itself. The repeated content also overlaps with that corpus. Among the 5-grams repeated by each model on the subset, 38.6% for LLaVA-Video-7B-Qwen2 and 20.7% for the Video-Only variant appear verbatim in LLaVA-Video-178K, compared with 1.9% to 22.8% in

five corpora that these models never saw, after matching vocabulary size. The models also show little tendency to terminate early. Their end-of-sequence probability averages 0.000 over the first 50 tokens and at the first repeated 5-gram, rising to only 0.019 over the last 50 tokens. Video-caption training therefore appears to induce a description length shaped more by annotation length than by the clip itself. A model trained on 1,055-word captions learns to generate at this scale, continues beyond the point where the clip has been described, and reuses patterns from the same corpus. This evidence remains observational, and a retraining study that varies caption length is a natural next step toward establishing the causal role of caption length.

Training Recipes Across Models. Repetition appears to be shaped more strongly by video-language alignment and post-training than by the base LLM alone. This pattern is clear among Qwen-based models, where RR ranges from 8–13 for InternVL3.5-8B and Qwen3-VL-8B-Instruct to 58–72 for LLaVA-Video-7B-Qwen2-Video-Only and Molmo2-8B, suggesting that the multimodal adaptation pipeline plays a major role in repetition.

Post-training also has a clear effect. LLaVA-NeXT-Video-7B-DPO consistently achieves lower RR than its non-DPO counterpart, suggesting that alignment objectives can mitigate repetition. Similarly, LLaVA-Video-7B-Qwen2 has much lower RR than its Video-Only variant. Since the two differ only in whether non-video data is included in the training mixture, removing that data shifts the model toward the repetitive style of its video captions, as shown by the audit above.

Output length is another contributing factor. Models that generate shorter responses on benign videos (*e.g.,* LLaVA-Video-7B-Qwen2, InternVL3.5-8B) are less likely to enter repetition, while more verbose models tend to exhibit longer repetition loops.

Models trained on caption-heavy or instruction-style datasets (*e.g.,* ShareGPT4Video) also tend to generate longer outputs and consequently more repetition. We do not observe a clear qualitative difference in the form of repetition. Both native multimodal and LLM-centric models produce similar phrase-level and sentence-level repetition.

C.6 Decoding-Time Mitigation

Decoding, Filtering, and Prompting. To test the robustness of the output repetition, we also conduct experiments by varying decoding settings, input filtering and prompts while keeping the rest of the evaluation pipeline unchanged. Figure 7 reports ShareGPT4Video at 32 frames on the full testbed. In Figure 7, T denotes temperature, p denotes top-p sampling [46], k denotes top-k sampling [47], r denotes repetition penalty [48], TRF denotes temporal redundancy filtering and P2, P3, and P4 denote the alternative prompts "Summarize the video content.", "What is happening in this video?", and "Explain the sequence of events.". The results show that repetition is largely stable across decoding choices other than the repetition penalty.

Temperature. Increasing temperature changes the randomness of the wording, but the overall repetition rate remains similar. This suggests that stochastic decoding changes local lexical variation more than it changes the underlying tendency to fall into repetitive loops. In other words, repetition is not simply an artifact of fully deterministic decoding. Our main experiments use T=0 because deterministic decoding is important for reproducibility and fair comparison across perturbation settings. This is also the default setting for most models.

Top-p and Top-k. The effect is limited and only partially helpful. Compared with plain T0.7 p0.9, adding top-k filtering reduces repetition to some extent, and using k50 alone reduces it further. A plausible explanation is that top-k truncation removes part of the low probability tail and can therefore suppress some unstable continuations. However, this improvement is not sufficient to eliminate repetition, and the overall trend remains the same after perturbations. Thus, these decoding heuristics may weaken the symptom, but they do not address the underlying cause.

Repetition Penalty. This is the only decoding strategy that consistently reduces repetition more substantially. For example, the repetition rate of original videos decreases from 82 to 56 with r1.1 and 27 with r1.2, and similar reductions appear in Add, Delete, and Replace. However, stronger repetition penalties distort normal token reuse, which is often necessary for fluent and faithful video description, and we quantify the resulting quality trade-off on the subset below.

Temporal Redundancy Filtering. We achieve this by removing highly similar or duplicate frames before feeding the video into the model. However, we observe that the repetition rate only decreases

slightly under this setting, and the overall behavior remains largely unchanged. This suggests that repetition is not solely caused by explicit frame-level redundancy in the input.

Prompt Variation. It also has only a minor effect. Although the absolute values fluctuate somewhat across P2, P3, and P4, repetition remains high overall and the relative trend across original videos and transformed videos is preserved. This indicates that the phenomenon is not highly sensitive to a particular prompt wording. Instead, it appears to be a more intrinsic behavior of current VideoLLMs when processing temporally redundant or temporally perturbed inputs.

Decoding Settings and Output Quality. We re-run several decoding strategies on LLaVA-Video-7B-Qwen2-Video-Only at 32 frames on the subset, add a stronger repetition penalty of 1.3, and score each setting jointly for repetition and output quality, as reported in Table 13. Following Figure 7, T denotes temperature, p denotes top-p sampling, and r denotes the repetition penalty, and Severe is the share of outputs that reach the token cap.

Table 13: Repetition and description quality under decoding settings.

Setting	RR (%)	Severe (%)	Words	CLIPScore	Distinct content words
T0 (greedy, default)	61	12	209	0.2323	72.3
T0.7 p0.9	60	20	227	0.2295	77.5
T0.7 p0.9 r1.1	50	8	212	0.2284	84.6
T0.7 p0.9 r1.2	36	0	224	0.2261	105.3
T0.7 p0.9 r1.3	10	0	225	0.2196	121.1

Stronger penalties suppress repetition more effectively. A mild penalty of r1.1 lowers RR from 60% to 50% relative to sampling without a penalty, while RR falls from 61% under greedy decoding to 36% at r1.2 and 10% at r1.3. The severe token-cap tail disappears entirely at r1.2. This reduction comes at a cost to visual grounding. CLIPScore decreases monotonically with penalty strength, by 1.7%, 2.7%, and 5.5% at r1.1, r1.2, and r1.3 relative to greedy decoding, indicating that stronger penalties reduce agreement with the video. Over the same range, the number of distinct content words increases by 17% to 68%, while output length grows by only 8%, so the captions do not simply become shorter or less informative. The penalty reduces repetition and increases lexical diversity at the expense of grounding. The same trade-off appears on LLaVA-Video-7B-Qwen2, one of the least repetitive models, where r1.3 lowers CLIPScore by 4.5% and increases output length by 48%. This suggests that the cost arises from the constraint itself rather than only from suppressing repetition. On one clip, greedy decoding repeats "in the black shirt" 13 times in 432 words, while r1.3 produces 417 words without repetition but introduces groceries and a phone that are absent from the video.

The penalty also changes what the model generates rather than how much it tends to generate. Across these settings, the output length of the Video-Only variant never decreases and remains between 209 and 227 words, while the stopping probability is unchanged at r1.3. The tendency to repeat is also model-dependent. Under the single deterministic protocol in Table 1, the 32-frame RR across evaluated models ranges from 7% to 82%, and the three models sharing a Qwen3-8B backbone, InternVL3.5-8B, Qwen3-VL-8B-Instruct, and Molmo2-8B, achieve 7%, 11%, and 72%.

We further evaluate standard accuracy under the same settings, with results reported in Table 14. The NExT-QA clips are the same as above, comprising 1,566 multiple-choice and 1,384 open-ended questions, while MVBench [10] adds 343 questions on different videos. Both multiple-choice sets are scored using a single option letter, and the open-ended answers are compared with references averaging 2.6 words. As the penalty increases, all three benchmarks produce identical answers question by question because one-letter or two-word responses cannot repeat. When we remove the single-word instruction and then request a detailed answer, the model still responds in 3 words and no answer repeats a 5-gram, indicating that long outputs arise primarily during video description. We therefore also evaluate the descriptions with a question-answering probe, where a text-only model answers NExT-QA questions using only the generated description (Caption QA). Caption QA drops from 64.1% without a penalty to 59.5% at r1.3, meaning that suppressing repetition reduces the useful information carried by the description by 9%. Standard accuracy alone would suggest that r1.3 is harmless, while the description-level metrics show otherwise. CLIPScore falls from 0.2295 to 0.2196, output length remains similar, and the number of distinct words rises by 56%. In practice, r1.2 removes 40% of the repetition for a 1.5% CLIPScore cost relative to sampling without a penalty, and no setting achieves low repetition while preserving the original grounding. Repetition penalties

therefore provide an effective symptom-level control with a measurable and monotonic cost in visual grounding, while VIDEOSTF provides the measurements needed to characterize this trade-off.

Table 14: Repetition, description quality, and benchmark accuracy (%) under decoding settings.

Setting	RR	CLIPScore	Caption QA	NExT-QA		MVBench
				Multiple choice	Open-ended	
T0 (greedy)	61	0.2323	63.9	83.3	16.5	58.9
T0.7 p0.9	60	0.2295	64.1	79.8	14.6	56.0
T0.7 p0.9 r1.1	50	0.2284	62.3	79.8	14.6	56.0
T0.7 p0.9 r1.2	36	0.2261	61.2	79.8	14.6	56.0
T0.7 p0.9 r1.3	10	0.2196	59.5	79.8	14.6	56.0

Test-Time Re-Querying. The obvious test-time remedies each have limitations. Input cleaning is ineffective because Section C.3 shows that redundancy does not explain repetition. As shown above, decoding-time intervention either leaves RR largely unchanged or reduces it at the cost of visual grounding. Table 2 further shows that reducing the frame count lowers RR only at 4 frames or fewer, which discards temporal content that the model is expected to use. We therefore consider a test-time remedy that preserves the weights, decoder, and frame budget, following the broader paradigm of defenses based on detection [49–51], correction [52], and perturbed-input response aggregation [53] without retraining the protected model. Specifically, RR is used as a detector, and only outputs flagged as repetitive are re-queried using a different temporal view of the same video, incurring no additional cost for benign outputs. On LLaVA-Video-7B-Qwen2-Video-Only at 32 frames on the subset, Table 15 reports the fraction of originally repetitive outputs that become non-repetitive under each view. Cascading all four views recovers 41% of repetitive outputs with an average of 3.18 queries per video across the subset. Most outputs remain repetitive under every view, making this a practical mitigation rather than a complete solution, consistent with a failure rooted in the model itself rather than in a particular temporal view.

Table 15: Recovery rate (%) of repetitive outputs under alternative temporal views.

Alternative view	Recovery rate
Shuffle	21
Reverse	15
2× speed re-sample	15
Delete 2 frames	12
Cascade over all four views	41

Overall, the repetition penalty is the only decoding, filtering, or prompting strategy we test that clearly lowers repetition, but it trades this gain for weaker visual grounding, which leaves it an incomplete remedy. Prior work on text-only LLMs [26, 30] also shows that repetition cannot be fully resolved by decoding strategies alone. Training-level changes, such as adjusting the composition and length of video-caption supervision as examined in Section C.5, or temporal-aware objectives that stabilize cross-frame interactions, are therefore a promising complement. These approaches require retraining models and are beyond the scope of this benchmark, so we leave them to future work.

C.7 Video Distributions

We further evaluate on longer and more complex videos by sampling from VideoMME [54], Ego4D [55], and CLEVRER [56], covering durations from 3 min to 1 h and scenarios requiring more complex temporal reasoning. As shown in Figure 11, repetition consistently persists, and the three high-repetition models, LLaVA-Video-7B-Qwen2-Video-Only, ShareGPT4Video, and Molmo2-8B, exhibit comparable repetition rates on long, reasoning-heavy, instructional, egocentric, and synthetic videos. While our VIDEOSTF testbed can readily incorporate these additional videos, it is important to note that the observed repetition is not specific to particular datasets or video distributions; rather, it generalizes across different settings and scales, reflecting a model-level phenomenon rather than a data-level artifact. The primary objective of this benchmark is to provide a controlled and reproducible framework for comparing models under a consistent protocol, rather than to achieve exhaustive coverage of all possible video domains and application scenarios.

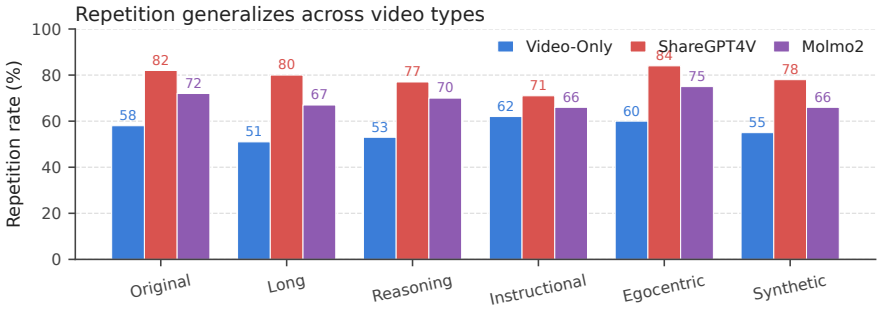

Figure 11: Repetition generalizes across video types, including long, reasoning, instructional, egocentric, and synthetic videos.

C.8 Reverse and Temporal Priors

Models are primarily trained on forward temporal sequences, so Add, Delete, and Replace largely preserve temporal structure, while Reverse fundamentally disrupts it. This breaks learned temporal priors and leads to qualitatively different generation behavior, making repetition less likely. This observation is consistent with prior findings (*e.g.,* Vid-SME [57]), which suggest that reversed temporal sequences deviate from the training distribution and are less effectively modeled by current video models. As a mitigation strategy, Reverse alone is deterministic and thus limited, and Section C.6 shows that cascading it with other temporal views recovers more repetitive outputs than any single view, although most outputs still repeat.

D Limitations and Broader Impacts

Limitations. VIDEOSTF is designed to measure output repetition in VideoLLMs rather than provide a complete remedy for the failure it reveals. Among the mitigations we examine, repetition penalties reduce repetition at the cost of weaker visual grounding, while re-querying repetitive outputs with alternative temporal views recovers only a subset of them. A remedy that suppresses repetition without degrading grounding therefore remains open. The benchmark also does not propose a new VideoLLM, and its conclusions are bounded by the 10 models and the temporal stressor families we evaluate. All of these models are open-weight, so proprietary models remain untested, and other forms of temporal change may expose further failure modes, although the consistency of our findings across model families and base LLMs suggests that the failure is widespread. Finally, the n-gram-based metrics are standard and widely adopted but capture surface-level repetition, so they would miss a loop that restates the same content in different words. Semantic-redundancy measures could therefore provide a complementary signal.

Broader Impacts. VIDEOSTF is designed to help researchers and practitioners diagnose generation failures, evaluate VideoLLMs more comprehensively, and develop more stable video-language systems. Our attack evaluation further shows that simple temporal transformations can trigger repetitive degeneration in black-box settings, potentially degrading service quality, increasing inference cost, or affecting system availability. We include this analysis to characterize the exploitability of repetition and motivate stronger robustness rather than to encourage attacks on deployed systems. We therefore recommend responsible evaluation practices, including clear documentation of experimental procedures and disclosure of potential risks. More broadly, improved evaluation of generation stability may help identify and mitigate failures before VideoLLMs are deployed in real-world applications.

Answer: The video begins with a view of a white wooden structure, possibly a deck or balcony, set against a backdrop of green trees and a clear blue sky. A person wearing a pink top is seen standing on the upper level of the structure, looking down. The scene transitions to show two individuals, one in black shorts and the other in a bikini top and jeans, standing on the lower level. They appear to be interacting with each other, possibly dancing or playing. The background remains consistent with the previous scene, featuring the same trees and sky. The narrative continues with the same two individuals now standing on the upper level of the structure. The person in black shorts is holding hands with the person in the bikini top and jeans, who is now wearing a red top. They seem to be engaged in a playful interaction, possibly dancing or preparing to jump into the water below. The background remains unchanged, maintaining the continuity of the setting. The video wraps up with a close-up view of the feet of the two individuals, still on the upper level of the structure. The person in black shorts is barefoot, while the person in the red top and jeans is also barefoot. The ground appears to be a light-colored concrete surface with some dark spots, possibly dirt or shadows. The background is not visible in this frame, focusing solely on the lower part of the individuals and the ground they are standing on.

Answer: The video features two men in a room with white walls and a wooden door. Both men are dressed in dark clothing, with one wearing a black jacket and the other in a dark shirt. The sequence begins with the man in the black jacket standing near the door, while the man in the dark shirt is positioned slightly to the right. They appear to be engaged in a conversation or interaction, with the man in the black jacket gesturing with his hands. As the video progresses, the man in the dark shirt moves closer to the man in the black jacket, and they continue their interaction. The man in the dark shirt then turns away from the man in the black jacket and walks towards the door, eventually exiting the frame. The man in the black jacket remains near the door, looking in the direction of the man who just left. The video concludes with the man in the black jacket still standing near the door, looking down at something in his hands.

(a) Examples of output repetition on LLaVA-Video-7B-Qwen2

Answer: The video opens in a dimly lit kitchen where a person in a black shirt and dark pants is seen near the sink, with various items like a blue cup and some papers scattered on the counter. The scene transitions to another individual holding a brown and white dog, who appears calm and relaxed. This person then sits at the kitchen counter, holding the dog close, while another person stands nearby, observing. The focus shifts to a close-up of the dog being held by the person in the black shirt, who is now sitting at the counter. The scene then shows two people in the kitchen, one holding the dog and the other standing nearby. The person holding the dog eventually stands up and walks away, leaving the other person seated at the counter. The narrative continues with the person in the black shirt sitting at the counter, holding the dog, while another person stands nearby. The scene transitions to a wider view of the kitchen, showing the person in the black shirt still holding the dog, and another person standing in the background. The person holding the dog then stands up and walks away, leaving the other person seated at the counter. The focus returns to the person in the black shirt, now holding the dog close, while another person stands nearby. The scene transitions to a wider view of the kitchen, showing the person in the black shirt still holding the dog, and another person standing nearby. The person holding the dog then stands up and walks away, leaving the other person seated at the counter. The video then shows a person in a black shirt and dark pants standing near the sink, holding a small object, possibly a phone or remote control. The scene transitions to a wider view of the kitchen, showing the same person standing near the refrigerator, which has various papers and notes attached to it. The person appears to be engaged in a conversation, gesturing with their hand. The video concludes with the person still standing near the refrigerator, continuing their conversation.

Answer: The video opens with a person riding a bicycle down a set of green concrete steps adorned with graffiti, dressed in a black shirt and shorts. The background features trees, a traffic light, and a few people sitting on the steps. The rider skillfully navigates the steps, maintaining balance and control. As the rider continues to descend, they reach the bottom of the steps and move towards a higher platform. The scene transitions to the rider standing at the top of the steps, preparing for another descent. The rider then begins to descend again, maintaining balance and control, and successfully reaches the bottom of the steps. The background remains consistent with trees, a traffic light, and a few people sitting on the steps. The rider moves towards a higher platform, continuing their journey. The rider is seen at the top of the steps once more, preparing for another descent. The rider descends the steps, maintaining balance and control, and successfully reaches the bottom. The background remains unchanged, featuring trees, a traffic light, and a few people sitting on the steps. The rider moves towards a higher platform, continuing their journey. The rider is seen at the top of the steps again, preparing for another descent. The rider descends the steps, maintaining balance and control, and successfully reaches the bottom. The background remains consistent with trees, a traffic light, and a few people sitting on the steps. The rider moves towards a higher platform, continuing their journey. The rider is seen at the top of the steps once more, preparing for another descent. The rider descends the steps, maintaining balance and control, and successfully reaches the bottom. The background remains unchanged, featuring trees, a traffic light, and a few people sitting on the steps. The rider moves towards a higher platform, continuing their journey. The rider is seen at the top of the steps again, preparing for another descent. The rider descends the steps, maintaining balance and control, and successfully reaches the bottom. The background remains consistent with trees, a traffic light, and a few people sitting on the steps. The rider moves towards a higher platform, continuing their journey. The rider is seen at the top of the steps once more, preparing for another descent. The rider descends the steps, maintaining balance and control, and successfully reaches the bottom. The background remains unchanged, featuring trees, a traffic light, and a few people sitting on the steps. The rider moves towards a higher platform, continuing their journey. The rider is seen at ...

(b) Examples of output repetition on LLaVA-Video-7B-Qwen2-Video-Only

Figure 12: Examples of repetitive outputs generated by *LLaVA-Video-7B-Qwen2* and *LLaVA-Video-7B-Qwen2-Video-Only*, with repeated phrases or sentences highlighted in different colors.

Answer: In the video, we see a group of people gathered in a room, each holding a sword. The room is dimly lit, with the only source of light coming from a window on the right side. The people are dressed in medieval clothing, complete with helmets and armor. They are standing in a line, facing each other, and appear to be engaged in a sword-fighting practice. The first person in the line is holding a sword in their right hand, while their left hand is holding a shield. The second person in the line is also holding a sword in their right hand, but their left hand is empty. The third person in the line is holding a sword in their right hand and a shield in their left hand. The fourth person in the line is holding a sword in their right hand and a shield in their left hand. The fifth person in the line is holding a sword in their right hand and a shield in their left hand. The people in the room are all focused on each other, their swords held high and ready for action. The sound of their footsteps echoes in the room, adding to the intensity of the scene. The video captures the essence of medieval sword-fighting, with the participants fully immersed in the practice.

Answer: In the video, we see two women walking down a sidewalk. They are dressed in casual attire and appear to be engaged in a conversation. As they walk, they pass by a white car parked on the side of the road. The car is parked on the side of the road, and it's a standard sedan with four doors. The women continue their walk, passing by a white truck parked on the side of the road. The truck is also a standard four-door model, and it's parked on the side of the road. The sidewalk they are walking on is made of concrete and appears to be in good condition. The road they are walking on is also paved and appears to be in good condition. The sky is clear and blue, indicating that it's a sunny day. The trees in the background are green and appear to be in full bloom, suggesting that it's spring or summer. Overall, the video captures a peaceful moment of two women walking down a sidewalk on a sunny day, passing by parked cars and enjoying the pleasant weather.

(a) Examples of output repetition on LLaVA-NeXT-Video-7B

Answer: In the video, we see a group of four men engaged in a lively conversation in an office setting. The room is well-lit, with a desk and a computer visible in the background. The men are casually dressed, suggesting a relaxed and informal atmosphere. One man is seated on the left side of the frame, holding a bottle in his hand. He appears to be actively participating in the conversation, contributing to the discussion with his gestures and expressions. In the center of the frame, another man is seated, his attention focused on the conversation. He seems to be listening intently, possibly waiting for his turn to speak or react to what's being said. On the right side of the frame, a third man is seated, also engaged in the conversation. He too appears to be actively participating, contributing to the discussion with his body language and facial expressions. The fourth man is standing, positioned slightly to the left of the frame. He is also actively participating in the conversation, contributing to the discussion with his gestures and expressions. The men's body language suggests a friendly and collaborative atmosphere, with everyone actively participating in the conversation. The setting, the lighting, and the casual attire of the men all contribute to the relaxed and informal nature of the scene.

Answer: In the heart of the video, a baby, adorned in a blue onesie, is the star of the show. The baby's eyes are filled with curiosity and wonder as they gaze up at the camera, their tiny hands reaching out towards it. The baby's mouth is open wide, as if caught mid-chew or perhaps in the middle of a delightful conversation. The baby's gaze is fixed on the camera, creating an intimate connection between the viewer and the baby. The baby's tiny hands are reaching out towards the camera, almost as if they are trying to touch the viewer's face. This action adds a sense of interaction and engagement to the video. The baby's mouth is open wide, perhaps in the middle of a delightful conversation or perhaps caught mid-chew. This action adds a sense of liveliness and energy to the video. The baby's eyes are filled with curiosity and wonder as they gaze up at the camera. This action adds a sense of innocence and wonder to the video. Overall, this video captures a beautiful moment of a baby's curiosity and wonder, creating an intimate connection between the viewer and the baby.

(b) Examples of output repetition on LLaVA-NeXT-Video-7B-DPO

Answer: The woman walking on the sidewalking a woman walking down the sidewalking a woman walking down the sidewalking a sidewalking a woman walking down the sidewalking a woman walking down the sidewalking woman.

Answer: In the video, a man is seen water skiing on a calm lake. He is wearing a life jacket and holding onto a rope that is attached to a boat. The man is skillfully gliding across the water, leaving a trail of white wake in his wake. The water skiing is smooth and steady. The maneuvering, maintaining his balance and control over the water skis. The maneuvering, and the boat's movement is steady and controlled. The maneuvering is steady and controlled. The maneuvering is steady and controlled. The maneuvering

Answer: The video begins with a split-screen showing two images of a blue toy horse, one standing and the other lying down. The text in Russian reads 'Удивительные Нахожение Нандельные Бутри' (Amazing Find: Nandele Butri), indicating a surprising discovery related to these toys. The scene transitions to a beach setting where a person stands facing the ocean, with hands holding two large seashells positioned to resemble wings. The text 'Julia S' appears on the screen, along with a red heart icon and a button labeled 'ПОДПИШИСЬ' (Subscribe) with an arrow pointing to the right, encouraging viewers to subscribe. The final scene of this segment shows two green plush toys, one resembling a cactus with a smiling face and the other with a sad face, placed on a flat surface. The narrative continues with a close-up of the same two green plush toys, maintaining their positions. The text 'Julia S' is visible on the screen. The scene then shifts to a close-up of a peeled orange, revealing its juicy segments, with the text 'Julia S' still present. Following this, a hand holding a small piece of fruit, possibly a seed or pit, is shown, with the text 'Julia S' remaining on the screen. The final scene of this segment features a white dice with a blue face, which appears to be broken, with the text 'Julia S' still visible. The video progresses with a close-up of the same white dice with a blue face, emphasizing its broken state. The scene transitions to a close-up of a damaged statue, with the text 'Julia S' on the screen. A person wearing blue gloves examines the statue closely, with the text 'Julia S' still visible. The next scene shows a split-screen view of two statues, one intact and the other with an X-rayed appearance revealing its internal structure, with the text 'Julia S' on both sides. The final scene of this segment features a close-up of a wooden log with a circular purple object embedded in it, with the text 'Julia S' visible on the screen. The video wraps up with a close-up of the wooden log featuring the circular purple object embedded in it. The text 'Julia S' is visible on the screen. The camera angle shifts slightly, providing different perspectives of the log and the embedded object, with the text 'Julia S' remaining on the screen throughout. The video concludes with this detailed view of the wooden log and the embedded object, maintaining the focus on the unique combination of natural and artificial elements.

Answer: The video begins with a weather forecast segment for Hurricane Irma, presented by a meteorologist in a white shirt and blue tie. The background features a map of the southeastern United States, highlighting the hurricane's path and intensity. The text on the screen reads 'Hurricane Irma' with details such as 'Max Winds: 130 MPH,' 'Moving NW at 8 MPH,' and 'Next advisory at 11 PM.' The time displayed is 8:00 AM. In the top right corner, there is a graphic showing the hurricane's projected path and intensity over time. The bottom right corner shows an aerial view of an airport with the text 'TAMPA INT'L' and 'MEMPHIS INT'L.' The meteorologist points to different areas on the map, explaining the hurricane's movement and impact. The video continues with the same weather forecast segment for Hurricane Irma. The meteorologist remains in the same attire, and the background still features the map of the southeastern United States. The text on the screen now reads 'Hurricane Irma' with details such as 'Max Winds: 130 MPH,' 'Moving NW at 8 MPH,' and 'Next advisory at 11 PM.' The time displayed is 8:00 AM. The top right corner still shows the graphic of the hurricane's projected path and intensity over time. The bottom right corner now shows an aerial view of an area with palm trees and a highway, indicating strong winds. The meteorologist continues to point to different areas on the map, explaining the hurricane's movement and impact. The video progresses with the same weather forecast segment for Hurricane Irma. The meteorologist remains in the same attire, and the background still features the map of the southeastern United States. The text on the screen now reads 'Hurricane Irma' with details such as 'Max Winds: 130 MPH,' 'Moving NW at 8 MPH,' and 'Next advisory at 11 PM.' The time displayed is 8:00 AM. The top right corner still shows the graphic of the hurricane's projected path and intensity over time. The bottom right corner now shows an aerial view of an area with palm trees and a highway, indicating strong winds. The meteorologist continues to point to different areas on the map, explaining the hurricane's movement and impact. The text at the bottom of the screen changes to show different locations and wind speeds, such as 'Monroe City at 155 MPH' and 'Columbia at 100 MPH.' The video wraps up with the same weather forecast segment for Hurricane Irma. The meteorologist remains in the same attire, and the background still features the map of the southeastern United States. The text on the screen now reads 'Hurricane Irma' with details such as 'Max Winds: 130 MPH,' 'Moving NW at 8 MPH,' and 'Next advisory at 11 PM.' The time displayed is 8:00 AM. The top right corner still shows the graphic of the hurricane's projected path and intensity over time. The bottom right corner now shows an aerial view of an area with palm trees and a highway, indicating strong winds. The meteorologist continues to point to different areas on the map, explaining the hurricane's movement and impact. The text at the bottom of the screen changes to show different locations and wind speeds, such as 'Memphis at 130 MPH' and 'Raleigh at 25 MPH.'

(a) Examples of output repetition on LLaVA-NeXT-Video-32B-Qwen

Answer: First, we see a man and a woman standing on a hill with a parachute. They are wearing helmets and seem to be preparing for a paragliding adventure. Next, we see them standing on a grassy hill with a parachute. They are still wearing their helmets and seem to be enjoying the view. Then, we see them standing on a hill with a parachute. They are wearing backpacks and seem to be ready for their adventure. In the following scene, we see them standing on a hill with a parachute. They are wearing helmets and seem to be taking in the view. Next, we see them standing on a hill with a parachute. They are wearing backpacks and seem to be preparing for their adventure. Then, we see them standing on a hill with a parachute. They are wearing helmets and seem to be enjoying the view. Finally, we see them standing on a hill with a parachute. They are wearing helmets and seem to be taking in the view. Overall, it seems like they are having a great time and enjoying the beautiful scenery.

Answer: In the video, there are two cats, a black and white cat and a brown and black cat. The black and white cat is standing over the brown and black cat, who is lying on their side. The black and white cat is licking the brown and black cat's face. The brown and black cat is lying still and appears to be sleeping. The black and white cat continues to lick the brown and black cat's face throughout the video.

(b) Examples of output repetition on VideoLLaMA2

Figure 14: Examples of repetitive outputs generated by *LLaVA-NeXT-Video-32B-Qwen* and *VideoLLaMA2*, with repeated phrases or sentences highlighted in different colors.

Answer: The video showcases a dance performance by a group of individuals on a stage, beginning with a single dancer in a white shirt and black pants performing a solo routine. This dancer executes a series of dynamic movements, including a leap, a turn, and a pose with one arm extended upwards, before transitioning into a more complex sequence involving multiple dancers. The group, dressed in white shirts and black pants, performs synchronized choreography that includes synchronized jumps, turns, and poses, with the dancers' movements becoming more intricate and coordinated as the performance progresses. As the routine unfolds, the dancers execute a series of synchronized jumps and turns, with their movements becoming more dynamic and expressive. The choreography includes a variety of poses and movements, with the dancers' arms and legs moving in unison, creating a visually striking performance. The lighting plays a significant role in highlighting the dancers' movements, with the stage bathed in a warm glow that accentuates the dancers' forms and adds depth to the performance. The performance reaches a climax with the dancers executing a series of synchronized jumps and turns, with their movements becoming more dynamic and expressive. The choreography includes a variety of poses and movements, with the dancers' arms and legs moving in unison, creating a visually striking performance. The lighting continues to play a significant role in highlighting the dancers' movements, with the stage bathed in a warm glow that accentuates the dancers' forms and adds depth to the performance. Throughout the video, the camera remains stationary, providing a consistent view of the performance. The focus is on the dancers and their movements, with the background remaining blurred and indistinct. The video captures the energy and precision of the dancers' movements, showcasing their skill and coordination.

Answer: The video captures a sequence of events in a serene outdoor setting, beginning with a tranquil scene of a stone staircase leading up to a lush green area, surrounded by trees and bushes. A person is seen descending the stairs, moving closer to the camera, suggesting a peaceful, natural environment. As the video progresses, the individual continues their descent, with the camera following their movement, maintaining a consistent focus on the staircase and the person's journey down it. The person's movement is steady, with the camera tracking their descent, capturing the details of the staircase and the surrounding greenery. The environment remains unchanged, with the camera's focus on the staircase and the person's movement down it. The person's descent is captured in a continuous, smooth motion, with the camera maintaining a steady pace to keep the individual in the frame. As the video continues, the person reaches the bottom of the staircase, and the camera's focus shifts slightly to capture the person's movement into the green space beyond the staircase. The person's journey is captured in a continuous, fluid motion, with the camera following their movement, maintaining a steady pace to keep the individual in the frame. The person's movement is captured in a continuous, fluid motion, with the camera following their movement, maintaining a steady pace to keep the individual in the frame. The person's journey is captured in a continuous, fluid motion, with the camera following their movement, maintaining a steady pace to keep the individual in the frame. The person's journey is captured in a continuous, fluid motion, with the camera following their movement, maintaining a steady pace to keep the individual in the frame. The person's journey is captured in a continuous, fluid motion, with the camera following their movement, maintaining a steady pace to keep the individual in the frame. The person's journey is captured in a continuous, fluid motion, with the camera following their movement, maintaining a steady pace to keep the individual in the frame. The person's journey is captured in a continuous, fluid motion, with the camera following their movement, maintaining a steady pace to keep the individual in the frame. The person's journey is captured in a continuous, fluid motion, with the camera following their movement, maintaining a steady pace ...

(a) Examples of output repetition on ShareGPT4Video

Answer: A baby wearing a white shirt and yellow shorts is sitting on a bed with a green blanket. The baby is holding a book and flipping through the pages. The baby looks at the book and then turns to look at someone off-camera. The baby then reaches out and touches the person's hand. The baby continues to look at the book and then turns to look at the person again. The baby then reaches out and touches the person's hand again. The baby then looks back at the book and continues to flip through the pages. The baby then looks up and smiles. The baby then reaches out and touches the person's hand one last time before the video ends.

Answer: A person is walking around a house with a dog. The dog is walking around the house and playing with another dog. The person is walking around the house and the dog is following them. The dog is playing with a ball and a toy. The person is sitting on a couch and the dog is lying on the floor. The dog is playing with a cat. The person is walking around the house and the dog is following them. The dog is playing with a ball and a toy. The person is sitting on a couch and the dog is lying on the floor. The dog is playing with a cat.

Answer: A man is in the kitchen washing dishes while another man is standing in the doorway. The man in the doorway is talking to the man washing dishes. The man in the doorway is holding a dog. The man washing dishes is laughing and wiping his face with a towel. The man in the doorway is smiling and looking at the dog. The man washing dishes puts the towel down and picks up a phone. He looks at the phone and smiles. The man in the doorway walks away.

(b) Examples of output repetition on InternVL3.5-8B

Figure 15: Examples of repetitive outputs generated by *ShareGPT4Video* and *InternVL3.5-8B*, with repeated phrases or sentences highlighted in different colors.

Answer: The video opens with a young woman standing on a wooden platform, her back to the camera. She is wearing a red bikini top and blue jeans, and she is smiling as she looks out over the water. The camera then pans to the right, showing a man in a red shirt and black shorts standing on the platform. He is also smiling and looking out over the water. The camera then pans to the left, showing a young man in a black shirt and black shorts standing on the platform. He is also smiling and looking out over the water. The camera then pans back to the right, showing the woman again. She is now walking towards the camera, still smiling. The camera then pans to the left, showing the man in the red shirt again. He is now walking towards the camera, still smiling. The camera then pans back to the right, showing the woman again. She is now standing still, looking at the camera. The camera then pans to the left, showing the man in the red shirt again. He is now standing still, looking at the camera. The camera then pans back to the right, showing the woman again. She is now walking away from the camera, still smiling. The camera then pans to the left, showing the man in the red shirt again. He is now walking away from the camera, still smiling. he camera then pans back to the right, showing the woman again. She is now standing still, looking at the camera. The camera then pans to the left, showing the man in the red shirt again. He is now standing still, looking at the camera. The camera then pans back to the right, showing the woman again. She is now walking away from the camera, still smiling. The camera then pans to the left, showing the man in the red shirt again. He is now walking away from the camera, still smiling. he camera then pans back to the right, showing the woman again. She is now standing still, looking at the camera. The camera then pans to the left, showing the man in the red shirt again. He is now standing still, looking at the camera. The camera then pans back to the right, showing the woman again. She is now walking away from the camera, still smiling. The camera then pans to the left, showing the man in the red shirt again. He is now walking away from the camera, still smiling. The camera then pans back to the right, showing the woman ...

Answer: The video opens with a close-up shot of a baby sitting in a blue high chair. The baby is wearing a white shirt and is holding a green cup with a straw. The baby is looking at the cup and appears to be drinking from it. The background is slightly blurred, but it appears to be a living room with a couch and a lamp. The camera then quickly pans to the right, showing a blurry image of a person's legs and feet. The person is wearing blue jeans and is sitting on the floor. The camera then pans back to the left, showing the baby again. The baby is now looking at the camera and smiling. The baby is still holding the green cup and appears to be happy. The video ends with the baby looking at the camera and smiling. The baby is still holding the green cup and appears to be happy. The background is still slightly blurred, but it appears to be the same living room as before.

(a) Examples of output repetition on Qwen3-VL-8B-Instruct

Answer: A dimly lit stage with a light brown wooden floor and a large white curtain backdrop hosts a breakdancing performance, the top of the scene fading into black and the bottom edge showing the silhouettes of audience heads. Two dancers in matching outfits—white shirts, black pants, and white shoes—open the act with synchronized moves. The dancer on the left, wearing a black hat, and the dancer on the right, wearing a green hat, both begin on their hands, then rise into handstands with legs extended straight up. The dancer on the left kicks his right leg outward while the dancer on the right kicks his left leg outward, their bodies aligned against the white curtain. The lighting is low and slightly blurry, with a small red light glowing in the upper right corner, and the audience remains mostly in shadow. A third dancer in a white shirt and black pants enters from the right, and a fourth dancer in a white shirt and black pants joins from the left, bringing the group to four. The two original dancers continue their handstand work, then transition into a cartwheel, while the newcomer on the right performs a cartwheel as well. The dancer on the left then drops into a split, and the dancer on the right mirrors the move with a split. The lighting shifts to a reddish hue, bathing the stage in a warm glow as the four continue their coordinated routine. The scene changes to a blue-lit stage, where the four dancers are now arranged in a line. The dancer on the far left wears a white shirt, black pants, and a black hat; the dancer next to him wears a white shirt, black pants, and a white hat; the third dancer wears a white shirt, black pants, and a black hat; and the dancer on the far right wears a white shirt, black pants, and a white hat. The dancer on the far left begins a dance move, and the others follow in sequence, maintaining the line formation as they perform together. The lighting shifts again to a greenish hue, and the dancers continue their synchronized choreography. The dancer on the far left now wears a white shirt, black pants, and a white hat; the dancer next to him wears a white shirt, black pants, and a white hat; the third dancer wears a white shirt, black pants, and a white hat; and the dancer on the far right wears a white shirt, black pants, and a white hat. The dancer on the far ...

Answer: A small child, about two years old, sits in a light blue inflatable kiddie pool outdoors on a sunny day, the pool's surface decorated with playful sea creatures—fish, starfish, and a turtle—while the surrounding ground is a mix of concrete and dirt with tree shadows. The child wears a blue swimsuit with a green front and a blue hat with a white stripe, and holds a green toy in the left hand and a yellow starfish in the right, looking down at the water. The pool is filled with water, and the child splashes and plays with the toys, including a green turtle and a yellow starfish. The child then sits in the center, wearing a blue hat with a white stripe and a blue-and-white swimsuit, holding a green toy in the left hand and a yellow toy in the right, looking to the left; the pool's sides show a fish and a starfish, and the ground around it is a mix of concrete and dirt with tree shadows. Later, the child sits in the pool wearing a blue hat with a white stripe and a blue-and-white swimsuit, holding a green toy in the left hand and a yellow toy in the right, looking down at the water; the pool's sides show a fish and a starfish, and the ground is a mix of concrete and dirt with tree shadows. Finally, the child sits in the pool wearing a blue hat with a white stripe and a blue-and-white swimsuit, holding a green toy in the left hand and a yellow toy in the right, looking down at the water; the pool's sides show a fish and a starfish, and the ground is a mix of concrete and dirt with tree shadows.

(b) Examples of output repetition on Molmo2-8B

Figure 16: Examples of repetitive outputs generated by *Qwen3-VL-8B-Instruct* and *Molmo2-8B*, with repeated phrases or sentences highlighted in different colors.

Original Video



Answer: The video begins with a serene view of a train station platform, where a few people are seen standing and waiting. The platform is bordered by yellow tactile paving, and the background features a grassy embankment, trees, and an overpass. A train approaches from the distance, gradually getting closer to the platform. As the train nears, two individuals are visible on the platform, one holding a bag and the other a phone. The train continues its approach, eventually arriving at the platform. The scene transitions to a close-up view of the train as it arrives at the platform. The train, white with blue accents, has a sign with Japanese characters 'かんあい' (Kanai) on its side. The camera captures the train's side, showing reflections of the surroundings, including trees and the overpass. Two individuals are seen standing near the edge of the platform, one in a green jacket and the other in black. The train comes to a stop, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The train doors open, and passengers begin to board. The scene concludes with a close-up of the train doors opening, and a person in a red jacket walking away from the train. The video then shows shows a close-up view of the train doors opening, with a person in a red jacket walking away from the train. The camera shifts to show the side of the train, which has a sign with Japanese characters 'かんあい' (Kanai) on it. The train is white with blue accents, and the reflection of the surroundings, including trees and the overpass, is visible on the train's windows. The camera moves along the length of the train, capturing more of the 'かんあい' (Kanai) sign and the reflections of the environment. Passengers are seen boarding the train, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The scene concludes with a close-up of the train doors closing, and a person in a black jacket walking away from the train. The final scene shows a close-up view of the train doors closing, with a person in a black jacket walking away from the train. The camera shifts to show the side of the train, which has a sign with Japanese characters 'かんあい' (Kanai) on ...

Add 1: Add Frame 3 before Frame 16



Answer: The video begins with a serene view of a train station platform, where a few people are seen standing and waiting. The platform is bordered by a yellow tactile paving strip, and the background features a grassy embankment, trees, and an overpass. A train approaches from the distance, gradually getting closer to the platform. As the train nears, two individuals are visible on the platform, one in a green jacket and the other in black clothing. The train continues its approach, eventually arriving at the platform. The scene transitions to a close-up view of the train as it arrives at the platform. The train, white with blue accents, has a sign with Japanese characters 'かんあい' (Kanai) on its side. The camera captures the train's side, showing reflections of the surroundings, including trees and the overpass. Two individuals are seen standing on the platform, one in a green jacket and the other in black clothing. The train comes to a stop, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The train doors open, and a person in a red jacket exits the train, followed by another individual in black clothing. The train doors close, and the train starts to move away from the platform. The video then shows a close-up view of the train as it moves away from the platform. The train, still white with blue accents, has the 'かんあい' (Kanai) sign on its side. The camera captures the train's side, showing reflections of the surroundings, including trees and the overpass. The train continues to move away, and the camera focuses on the back of the train, highlighting the 'かんあい' (Kanai) sign. The train moves further away, and the camera captures the platform and the surrounding area, including the grassy embankment and trees. The train is no longer in view, and the platform remains empty. The final scene shows a close-up view of the train as it moves away from the platform. The train, still white with blue accents, has the 'かんあい' (Kanai) sign on its side. The camera captures the train's side, showing reflections of the surroundings, including trees and the overpass. The train continues to move away, and the camera focuses on the back of the train, highlighting the 'かんあい' ...

Add 2: Add Frames 3 and 1 before Frames 10 and 1



Answer: The video begins with a serene view of a train station platform, where a few people are seen standing and waiting. The platform is bordered by yellow tactile paving, and the background features a grassy embankment, trees, and an overpass. A train approaches from the distance, gradually getting closer to the platform. As the train nears, two individuals are visible on the platform, one holding a bag and the other talking on the phone. The train continues its approach, almost reaching the platform. The scene transitions to a close-up view of the train as it approaches the platform. The train, white with blue accents, has a sign with Japanese characters 'かんあい' (Kanai) on its side. Two individuals are seen standing on the platform, one in a teal jacket and black pants, and the other in a black outfit. As the train gets closer, the individuals move slightly to the side. The train passes by the platform, revealing more of its length and the station's surroundings, including the grassy embankment and trees. The train continues to move past the platform, showing more of its side and the station's infrastructure. Next, the video shows a close-up view of the side of the train, which is white with blue accents and has a sign with Japanese characters 'かんあい' (Kanai) on it. The train moves past the platform, revealing more of its length and the station's surroundings, including the grassy embankment and trees. The camera follows the train as it moves away from the platform, showing more of the station's infrastructure, such as the overpass and the surrounding area. The train continues to move away, with the camera capturing the side of the train and the station's environment. The video wraps up with a iew of the side of the train, which is white with blue accents and has a sign with Japanese characters 'かんあい' (Kanai) on it. The train moves past the platform, revealing more of its length and the station's surroundings, including the grassy embankment and trees. The camera follows the train as it moves away from the platform, showing more of the station's infrastructure, such as the overpass and the surrounding area. The train continues to move away, with the camera capturing the side of the train and the station's environment.

Figure 17: Example video under *temporal frame insertion*. We show the sampled frames and corresponding outputs from LLaVA-Video-7B-Qwen2-Video-Only at 16 frames for the original video, as well as the videos after transformations of *Add 1* frame and *Add 2* frames, with repeated phrases or sentences highlighted in different colors.

Original Video



Answer: The video begins with a serene view of a train station platform, where a few people are seen standing and waiting. The platform is bordered by yellow tactile paving, and the background features a grassy embankment, trees, and an overpass. A train approaches from the distance, gradually getting closer to the platform. As the train nears, two individuals are visible on the platform, one holding a bag and the other a phone. The train continues its approach, eventually arriving at the platform. The scene transitions to a close-up view of the train as it arrives at the platform. The train, white with blue accents, has a sign with Japanese characters 'かんあい' (Kanai) on its side. The camera captures the train's side, showing reflections of the surroundings, including trees and the overpass. Two individuals are seen standing near the edge of the platform, one in a green jacket and the other in black. The train comes to a stop, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The train doors open, and passengers begin to board. The scene concludes with a close-up of the train doors opening, and a person in a red jacket walking away from the train. The video then shows shows a close-up view of the train doors opening, with a person in a red jacket walking away from the train. The camera shifts to show the side of the train, which has a sign with Japanese characters 'かんあい' (Kanai) on it. The train is white with blue accents, and the reflection of the surroundings, including trees and the overpass, is visible on the train's windows. The camera moves along the length of the train, capturing more of the 'かんあい' (Kanai) sign and the reflections of the environment. Passengers are seen boarding the train, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The scene concludes with a close-up of the train doors closing, and a person in a black jacket walking away from the train. The final scene shows a close-up view of the train doors closing, with a person in a black jacket walking away from the train. The camera shifts to show the side of the train, which has a sign with Japanese characters 'かんあい' (Kanai) on ...

Delete 1: Delete Frame 6



Answer: The video begins with a serene view of a train station platform, where a few people are seen waiting. The platform is bordered by yellow tactile paving, and the background features a grassy embankment, trees, and an overpass. A train approaches from the distance, gradually getting closer to the platform. As it nears, two individuals are visible on the platform, one in a green jacket and the other in dark clothing. The train continues its approach until it reaches the platform, and the doors open, ready for passengers to board. The scene transitions to a close-up view of the train as it arrives at the platform. The train's blue and white exterior is prominently displayed, with large windows reflecting the surroundings. The platform remains bordered by yellow tactile paving, and the background still shows the grassy embankment and trees. The train comes to a stop, and the doors open, revealing a person in a red jacket standing at the door. Another individual in dark clothing walks along the platform. The train doors then close, and the train starts to move away from the platform. Next, the video shows a close-up view of the train as it moves away from the platform. The train's blue and white exterior is again prominently displayed, with large windows reflecting the surroundings. The platform is bordered by yellow tactile paving, and the background features a grassy embankment, trees, and an overpass. The train continues to move away, and the camera follows its movement, capturing the reflection of the surroundings on the train's windows. The train passes by a person in a red jacket standing at the door. The train then slows down and comes to a stop, with the doors opening to reveal a person in dark clothing walking along the platform. The train doors close once more, and the train starts to move away from the platform. The video wraps up with a view of the train moving away from the platform. The train's blue and white exterior is prominently displayed, with large windows reflecting the surroundings. The platform is bordered by yellow tactile paving, and the background features a grassy embankment, trees, and an overpass. The train continues to move away, and the camera follows its movement, capturing the reflection of the surroundings on the train's windows. The train passes by a person in a red jacket standing at the door. The train then slows down and comes to a stop, with the doors opening to reveal a person in dark clothing walking along the platform.

Delete 2: Delete Frames 6 and 10



Answer: The video begins with a serene view of a train station platform, where a few people are waiting. The platform is bordered by a yellow tactile paving strip, and the background features a grassy embankment, trees, and an overpass. A train approaches from the distance, gradually getting closer to the platform. As it nears, two individuals are seen standing on the platform, one in a green jacket and the other in black clothing. The train continues its approach until it reaches the platform, and the doors open, ready for passengers to board. The scene transitions to a close-up view of the train as it arrives at the platform. The train's blue and white exterior prominently displays Japanese characters, 'ふじや' (Fujiya), which translates to 'Fujiya' in English. The train doors open, and a person in a red jacket steps off, followed by another individual in a dark jacket. The camera captures the movement of the train and the actions of the passengers as they disembark. The background remains consistent with the previous scene, featuring the same grassy embankment, trees, and overpass. The video continues with a view of the train's side, again showing the Japanese characters 'ふじや' (Fujiya). The train doors open, and a person in a red jacket steps off, followed by another individual in a dark jacket. The camera then shifts to show more passengers disembarking, including a person in a green jacket and another in black clothing. The background remains unchanged, maintaining the grassy embankment, trees, and overpass. The scene concludes with a close-up of the train's side, highlighting the 'ふじや' (Fujiya) logo. The final scene shows a close-up view of the train's side, once again displaying the Japanese characters 'ふじや' (Fujiya).The train doors open, and a person in a red jacket steps off, followed by another individual in a dark jacket. The camera captures the movement of the train and the actions of the passengers as they disembark. The background remains consistent with the previous scenes, featuring the same grassy embankment, trees, and overpass. The video wraps up with a close-up of the train's side, emphasizing the 'ふじや' (Fujiya) logo.

Figure 18: Example video under *temporal frame deletion*. We show the sampled frames and corresponding outputs from LLaVA-Video-7B-Qwen2-Video-Only at 16 frames for the original video, as well as the videos after transformations of *Delete 1* frame and *Delete 2* frames, with repeated phrases or sentences highlighted in different colors.

Original Video



Answer: The video begins with a serene view of a train station platform, where a few people are seen standing and waiting. The platform is bordered by yellow tactile paving, and the background features a grassy embankment, trees, and an overpass. A train approaches from the distance, gradually getting closer to the platform. As the train nears, two individuals are visible on the platform, one holding a bag and the other a phone. The train continues its approach, eventually arriving at the platform. The scene transitions to a close-up view of the train as it arrives at the platform. The train, white with blue accents, has a sign with Japanese characters 'かんあい' (Kanai) on its side. The camera captures the train's side, showing reflections of the surroundings, including trees and the overpass. Two individuals are seen standing near the edge of the platform, one in a green jacket and the other in black. The train comes to a stop, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The train doors open, and passengers begin to board. The scene concludes with a close-up of the train doors opening, and a person in a red jacket walking away from the train. The video then shows shows a close-up view of the train doors opening, with a person in a red jacket walking away from the train. The camera shifts to show the side of the train, which has a sign with Japanese characters 'かんあい' (Kanai) on it. The train is white with blue accents, and the reflection of the surroundings, including trees and the overpass, is visible on the train's windows. The camera moves along the length of the train, capturing more of the 'かんあい' (Kanai) sign and the reflections of the environment. Passengers are seen boarding the train, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The scene concludes with a close-up of the train doors closing, and a person in a black jacket walking away from the train. The final scene shows a close-up view of the train doors closing, with a person in a black jacket walking away from the train. The camera shifts to show the side of the train, which has a sign with Japanese characters 'かんあい' (Kanai) on ...

Replace 1: Replace Frame 13 with Frame 1



Answer: The video begins with a close-up view of a train moving along the tracks, reflecting the surroundings in its large windows. The train is white with blue accents and has Japanese text on it. As the train continues to move, the camera angle shifts to show more of the train's side and the platform, where a few people are standing. The scene transitions to a wider view of the train station, showing the train approaching the platform. The background includes trees, a bridge, and some construction equipment. The train comes to a stop at the platform, and the camera captures the front of the train as it stops. The scene then shifts to a view of the train station platform, where two people are standing near the edge, looking towards the tracks. The background shows a bridge overhead and some greenery. The camera then focuses on the train tracks, where a train is seen approaching from the distance. The train, white with blue accents and featuring Japanese text, moves closer to the platform. The camera captures the train's side as it passes by, showing reflections of the surroundings in the windows. The train continues to move along the tracks, and the camera follows it, capturing the front of the train as it moves away from the platform. Next, the video shows a view of the train station platform, where a person in a red jacket is walking towards the edge of the platform. The background includes a bridge overhead and some greenery. The camera then focuses on the train tracks, where a train is seen approaching from the distance. The train, white with blue accents and featuring Japanese text, moves closer to the platform. The camera captures the train's side as it passes by, showing reflections of the surroundings in the windows. The train continues to move along the tracks, and the camera follows it, capturing the front of the train as it moves away from the platform. The scene concludes with a close-up view of the train's side, showing more details of the Japanese text and the reflections in the windows. The final scene shows a close-up view of the train's side, reflecting the surroundings in its large windows. The train is white with blue accents and has Japanese text on it. The camera then shifts to show the front of the train as it moves along the tracks. A person in a red jacket is seen walking towards the edge of the platform. The background includes a bridge overhead and some greenery. The camera captures the train's side again, showing reflections of the surroundings in the ...

Replace 2: Replace Frames 10 and 15 with Frames 2 and 4



Answer: The video begins with a view of a train station platform, where a blue and white train is seen arriving. The platform is bordered by yellow tactile paving, and several signs, including one indicating a pedestrian crossing, are visible. As the train approaches, it passes by the camera, revealing more of its blue exterior. The scene transitions to a close-up of the train's side, showing reflections of trees and the overpass above the station. The train continues to move along the tracks, passing by the camera again, and the platform remains visible in the background. The narrative develops through a close-up view of the blue and white train as it moves along the tracks. The train's side reflects the surrounding environment, including trees and an overpass. The camera follows the train as it passes by, capturing more of its blue exterior and the reflections on the windows. The platform with yellow tactile paving is still visible in the background. The train continues its journey, moving away from the camera, and the scene concludes with the train moving further along the tracks, with the platform and surrounding area still in view. The video wraps up with another view of the train station platform, similar to the previous scenes. A blue and white train is seen arriving, with the platform bordered by yellow tactile paving and several signs, including one indicating a pedestrian crossing. As the train approaches, it passes by the camera, revealing more of its blue exterior. The camera follows the train as it moves along the tracks, capturing more of its blue exterior and the reflections on the windows. The platform remains visible in the background. The train continues its journey, moving away from the camera, and the scene concludes with the train moving further along the tracks, with the platform and surrounding area still in view.

Figure 19: Example video under *temporal frame replacement*. We show the sampled frames and corresponding outputs from LLaVA-Video-7B-Qwen2-Video-Only at 16 frames for the original video, as well as the videos after transformations of *Replace 1* frame and *Replace 2* frames, with repeated phrases or sentences highlighted in different colors.

Original Video



Answer: The video begins with a serene view of a train station platform, where a few people are seen standing and waiting. The platform is bordered by yellow tactile paving, and the background features a grassy embankment, trees, and an overpass. A train approaches from the distance, gradually getting closer to the platform. As the train nears, two individuals are visible on the platform, one holding a bag and the other a phone. The train continues its approach, eventually arriving at the platform. The scene transitions to a close-up view of the train as it arrives at the platform. The train, white with blue accents, has a sign with Japanese characters 'かんあい' (Kanai) on its side. The camera captures the train's side, showing reflections of the surroundings, including trees and the overpass. Two individuals are seen standing near the edge of the platform, one in a green jacket and the other in black. The train comes to a stop, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The train doors open, and passengers begin to board. The scene concludes with a close-up of the train doors opening, and a person in a red jacket walking away from the train. The video then shows shows a close-up view of the train doors opening, with a person in a red jacket walking away from the train. The camera shifts to show the side of the train, which has a sign with Japanese characters 'かんあい' (Kanai) on it. The train is white with blue accents, and the reflection of the surroundings, including trees and the overpass, is visible on the train's windows. The camera moves along the length of the train, capturing more of the 'かんあい' (Kanai) sign and the reflections of the environment. Passengers are seen boarding the train, and the camera focuses on the front of the train, highlighting the 'かんあい' (Kanai) sign. The scene concludes with a close-up of the train doors closing, and a person in a black jacket walking away from the train. The final scene shows a close-up view of the train doors closing, with a person in a black jacket walking away from the train. The camera shifts to show the side of the train, which has a sign with Japanese characters 'かんあい' (Kanai) on ...

Reverse



Answer: The video begins with a person walking past a blue train adorned with Japanese text and graphics, including the word 'ふじや' (Fujiya), which translates to 'Fujiya' in English. The scene transitions to a train station platform where a person in a red jacket is seen walking towards the train. The camera then captures the side of the train as it moves along the tracks, with reflections of the surroundings visible on its windows. The train continues its journey through the station, passing by the platform. The narrative develops as the train, still marked with 'ふじや' (Fujiya), moves along the tracks. The camera follows the train from the side, capturing reflections of the surroundings, including trees and a bridge overhead. The train passes by the platform, and the scene shifts to a view from the platform, showing people waiting and watching the train as it departs. The train continues its journey through the station, moving away from the platform. The story progresses with a view from the platform, showing the blue train moving away. The camera captures the train from the side, with reflections of the surroundings, including trees and a bridge overhead. The train continues its journey through the station, moving further away from the platform. The scene then shifts to a view from the platform, showing people waiting and watching the train as it departs. The train continues its journey through the station, moving away from the platform. The video wraps up with a view of the train moving away from the platform, with reflections of the surroundings, including trees and a bridge overhead. The train continues its journey through the station, moving further away from the platform. The scene then shifts to a view from the platform, showing people waiting and watching the train as it departs. The train continues its journey through the station, moving away from the platform.

Shuffle



Answer: The video begins with a close-up of a blue train adorned with Japanese text and graphics, including the word 'あすかん' (Asukan), which translates to 'morning' in English. The scene transitions to a wider view of the train station platform, where people are seen waiting. The train is stationary, and the background features a bridge and some trees. The camera then shifts to show the side of the train, revealing more of the station and the surrounding area, including a sign with Japanese text. The train starts moving away from the station, and the camera captures it from different angles as it departs. Next, the video shows a white train with blue accents moving along the tracks. The camera follows the train as it passes by a platform where two people are standing, one in a green jacket and the other in a black outfit. The background includes a bridge and some trees. The scene transitions to a close-up of the train's side, showing more details of the station and the surrounding area, including a sign with Japanese text. The train continues to move along the tracks, and the camera captures it from different angles as it progresses. The video then shows a white train with blue accents moving along the tracks. The camera follows the train as it passes by a platform where a person in a red jacket is seen walking. The background includes a bridge and some trees. The scene transitions to a close-up of the train's side, showing more details of the station and the surrounding area, including a sign with Japanese text. The train continues to move along the tracks, and the camera captures it from different angles as it progresses. The video concludes with the train moving further away from the station, with the camera capturing it from a distance. Finally, the video shows a blue train with Japanese text and graphics, including the word 'あすかん' (Asukan), moving along the tracks. The camera follows the train as it passes by a platform where a person in a black outfit is seen walking. The background includes a bridge and some trees. The scene transitions to a close-up of the train's side, showing more details of the station and the surrounding area, including a sign with Japanese text. The train continues to move along the tracks, and the camera captures it from different angles as it progresses. The video concludes with the train moving further away from the station, with the camera capturing it from a distance.

Figure 20: Example video under *temporal reversal* and *temporal shuffling*. We show the sampled frames and corresponding outputs from LLaVA-Video-7B-Qwen2-Video-Only at 16 frames for the original video, as well as the videos after transformations of *Reverse* and *Shuffle*, with repeated phrases or sentences highlighted in different colors.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are consistent throughout the whole paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not include theories.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the link for our source code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the data we use is publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experiment setup is provided in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Our work is a benchmark and measurement study rather than a stochastic training-based method. The reported results are deterministic evaluations under certain settings, so statistical significance analysis or error bars are less applicable in this context. Moreover, the observed repetition trends are consistent across models and perturbation settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide it in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss positive and negative societal impacts, including the potential misuse of the black-box attack, in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite their work in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new assets are released.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: There is no crowdsourcing or research with human subjects in this paper.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: There is no crowdsourcing or research with human subjects in this paper.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.