GenState-AI: State-Aware Dataset for Text-to-Video Retrieval on AI-Generated Videos

Minghan Li[*,†], **Tongna Chen**[*], **Tianrui Lv**[*]
Yishuai Zhang, **Suchao An**, **Guodong Zhou**
School of Computer Science and Technology, Soochow University, Suzhou, China
[*]The first three authors contributed equally. [†]Corresponding author.
mhli@suda.edu.cn

Abstract

Existing text-to-video retrieval benchmarks often permit matching from static appearance, making it difficult to determine whether models capture state changes and final outcomes. We introduce GenState-AI[1], a benchmark comprising 956 AI-generated videos and 195 verified state-change queries across Toy, Cartoon, Home, and Manipulation domains. Its construction uses prompt-controlled variants that target either conflicting temporal/state outcomes or semantic substitutions. Following human verification, all valid videos contribute to subset-level retrieval pools, while complete query-specific triplets support diagnostic hard-negative ordering. We evaluate VCF-Lik, Qwen3-VL, and LanguageBind using both standard Recall@K and triplet-based consistency metrics. Qwen3-VL (8B) achieves the best macro-average R@1 and triplet consistency, while LanguageBind obtains the highest semantic discrimination accuracy. However, stronger retrieval does not consistently yield correct hard-negative ordering, and Qwen3-VL and LanguageBind remain weak on Manipulation triplets, where VCF-Lik performs substantially better. GenState-AI therefore provides a controlled diagnostic testbed for distinguishing global retrieval quality from fine-grained temporal and semantic grounding in AI-generated videos.

1 Introduction

Text-to-video retrieval aims to find the most relevant video given a natural language query, or vice versa, from large-scale collections (Miech et al., 2019; Gabeur et al., 2020; Luo et al., 2022; Ma et al., 2022; Reddy et al., 2025; Wang et al., 2024; Zhang et al., 2025). Recent work has shown that multi-modal large language models (MLLMs) can significantly improve retrieval, especially for long or compositional queries (Lin et al., 2025; Chen et al., 2025). A recent highlight is BLiM (Ko et al., 2025), which trains an MLLM to estimate bidirectional likelihoods (text given video and video features given text) and mitigates candidate prior bias using Candidate Prior Normalization (CPN), achieving strong performance on several popular real-world text–video benchmarks.

In parallel, modern video generative models such as Wan2.2-TI2V-5B (Wan et al., 2025), Step-Video-TI2V (Huang et al., 2025) and Make-A-Video (Singer et al., 2023), among others (Ha-Cohen et al., 2024; Ma et al., 2024; Zhou et al., 2022; Hong et al., 2022; Zhao et al., 2025; Liu et al., 2024a), can synthesize high-resolution, temporally coherent videos conditioned on text or image prompts.

These models make it possible to construct synthetic benchmarks that go beyond the biases and constraints of Internet videos, for example by enforcing controlled state transitions or non-physical but semantically precise scenarios.

However, most existing retrieval benchmarks, and thus most retrieval methods implicitly assume that (1) a single frame conveys most of the semantics, and (2) videos follow natural visual statistics. Consequently, negative samples in real-world datasets often differ from positives in many uncontrolled ways, allowing models to succeed through appearance matching rather than genuine temporal reasoning. Under these conditions, models need not explicitly track how visual states change over time.

This limitation raises a fundamental question:

> How can we reliably evaluate whether retrieval models are genuinely state- and temporal-aware, especially in the context of controllable AI-generated videos?

To address this gap, we construct prompt-controlled hard negatives that target specific state

[1]Dataset: https://huggingface.co/datasets/Mh-Li/genstate-ai

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2435–2448
November 6-10, 2026 ©2026 Association for Computational Linguistics

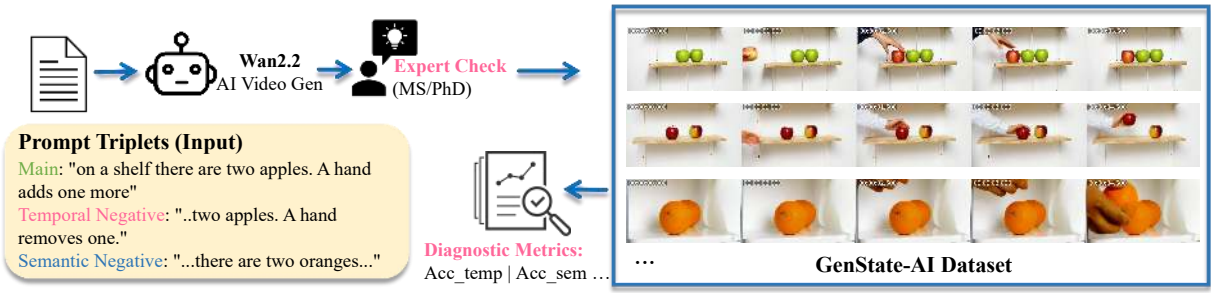


Figure 1: Overview of the GenState-AI construction and evaluation pipeline, including prompt-triplet construction, AI-based video generation, human verification and filtering, and retrieval evaluation.

or semantic conflicts, while acknowledging that independently generated videos may retain low-level visual variations.

GenState-AI We introduce GenState-AI, an AI-generated text-to-video retrieval benchmark explicitly designed around controlled state transitions (the process is shown in Fig. 1). We leverage the open-source Wan2.2-TI2V-5B hybrid text-image-to-video model (Wan et al., 2025), which can generate 720p videos at 24 fps, to procedurally create a large collection of short clips. Each clip is defined by a simple, human-readable prompt describing: (1) a small set of tabletop objects, (2) an initial configuration, and (3) a final configuration after a short motion or state change.

Crucially, we also construct hard negative prompts through minimal edits that keep the intended scene, objects, and configuration fixed at the prompt level while targeting a conflicting final state. For example, the teddy bear may move slightly but ultimately fails to cross to the other side of the mug. Such pairs create retrieval scenarios where neither global appearance nor superficial motion cues are sufficient; models must accurately infer the resulting state of the scene to succeed. Because the clips are independently generated, low-level visual appearance is not perfectly controlled; we therefore rely on human verification and diagnostic triplet analyses to reduce confounding effects.

We conduct a systematic evaluation of MLLM-based text-to-video retrieval on the proposed GenState-AI benchmark using four retrieval systems from three representative model families: (i) a likelihood-scoring T2V baseline built upon **VideoChat-Flash-Qwen2-7B_res448**, (ii) **Qwen3-VL** scale-matched embedding–reranking baselines with 2B and 8B embedding/reranker configurations, and (iii) a **LanguageBind** video–text embedding baseline. Despite their architectural differences, these baselines exhibit consistent failure patterns in synthetic, state-centric settings. When evaluated on prompt-controlled AI-generated videos, they often rank candidates with globally "on-topic" appearance or temporally plausible but end-state-incorrect trajectories ahead of the true match, indicating insufficient grounding to decisive end-state evidence. Scaling the Qwen3-VL configuration from 2B to 8B, with scale-matched embedding and reranking models, improves the macro-average R@1 from 0.539 to 0.581, while Acc_{tri} increases only modestly from 0.693 to 0.721, showing that GenState-AI is sensitive to model capability while still exposing residual temporal-ordering failures. However, this improvement is not uniform: Qwen3-VL (8B) still struggles on *Manipulation* (R@1: 0.304; Acc_{tri}: 0.524), while VCF-Lik achieves much stronger triplet consistency on the same subset (0.810).

The triplet design exposes a consistent asymmetry across baselines: temporal hard negatives are harder to reject than semantic substitutions, even when the correct video ranks near the top. This gap suggests that strong Recall@K does not necessarily imply reliable state-transition grounding, as models may favor globally "on-topic" or temporally plausible clips over evidence of the correct end state.

We further provide subset-level and triplet-order analyses to characterize temporal versus semantic confusions, highlighting open challenges for state- and time-sensitive text-to-video retrieval.

2 Related Work

Text-to-video retrieval has traditionally relied on dual encoders that align video and text representations through contrastive learning (Sun et al., 2019; Miech et al., 2019; Gabeur et al., 2020; Luo et al., 2022). More recent systems incorporate stronger

multimodal encoders, rerankers, or generative likelihood scoring (Hur et al., 2025; Chen et al., 2023; Girdhar et al., 2023; Li et al., 2026c,b; Zhu et al., 2024). For long-document text retrieval (Li et al., 2026a), prior work has explored BERT-based dense intra-ranking and contextualized late interaction (Li and Gaussier, 2022), as well as key-block selection with local pre-ranking (Li et al., 2023). More recent LLM-based rerankers select and compact query-relevant evidence before scoring (Li et al., 2024). These model-centric advances improve retrieval on common datasets, but standard benchmarks such as MSR-VTT, DiDeMo, ActivityNet-Captions, and VATEX (Xu et al., 2016; Anne Hendricks et al., 2017; Krishna et al., 2017; Wang et al., 2019) are dominated by real-world footage where global appearance often suffices. In text-only dense retrieval, cross-domain generalization has also been studied through self-supervised pseudo-relevance labeling (Li and Gaussier, 2024).

Recent diagnostic benchmarks probe temporal and fine-grained video understanding more directly. RTime (Du et al., 2024) constructs time-reversed hard negatives, CaReBench (Xu et al., 2024) separates spatial and temporal annotations, and TempCompass (Liu et al., 2024b) evaluates event order and attribute changes. Synthetic and generated-video datasets have also emerged (Liu et al., 2023; Chen and Krähenbühl, 2022; Guo et al., 2024; Wan et al., 2025), but they primarily target physics, controllability, or generation quality rather than retrieval. GenState-AI is complementary: it uses AI-generated clips and paired temporal/semantic hard negatives to diagnose whether retrieval models can distinguish the intended state outcome from visually plausible alternatives.

3 GenState-AI Dataset

3.1 Design Principles

GenState-AI is built around three principles:

1. **State-centric semantics.** Each video is defined by a clear and observable change in state. One category of state transitions arises from the actor manipulating objects, such as moving a target object to alter its position or modifying the quantity of objects by adding or removing instances. Another category is driven by the actor's own positional change, where the actor's movement within the scene constitutes an independent form of state transition. Together, these two types of state changes enable the dataset to systematically capture action-induced state transformations.

2. **Minimalist prompts.** We employ concise and unambiguous prompts that focus solely on the target state transition. Avoiding long or stylistically complex descriptions prevents semantic drift and ensures that both the model and the evaluation concentrate on the intended change of state.

3. **Hard negatives by state misalignment.**

   For each main query, we construct hard-negative prompts through minimal prompt edits that target either a conflicting end-state or a semantic substitution while preserving the intended scene and object configuration at the prompt level. These hard negatives fall into two categories. First, **semantic hard negatives** modify the target entity: for example, in tasks where an actor is supposed to move a specific object, the hard negative instead moves a different object; in scenarios where the actor's own movement is defined relative to a reference object, we replace the reference to induce a semantic mismatch. Second, **temporal hard negatives** alter the action dynamics themselves, such as keeping the object stationary when the positive sample requires movement, or performing only a partial motion that fails to reach the final state. Because the clips are generated independently from text-only prompts, low-level visual appearance is not perfectly controlled: background layout, lighting, object texture, pose, and generation artifacts may vary across videos in a triplet. We therefore rely on human verification and triplet-level diagnostics to reduce confounding effects and to ensure that each retained hard negative remains an interpretable contrast to the main query. These prompt-controlled hard negatives encourage models to rely on state reasoning rather than superficial topic cues, enabling a more stringent and fine-grained assessment of state understanding.

3.2 Video Generation with Wan2.2-TI2V-5B

We use the Wan2.2-TI2V-5B model (Wan et al., 2025), an open-source text-image-to-video generator that supports 1280×704 resolution at 24 fps with a 5B-parameter hybrid architecture.

Table 1: Number of query descriptions per video category in our GenState-AI dataset, along with the average token length, computed using the Qwen2.5-1.5B tokenizer.

Category	Number of Queries	Avg. Tokens
Home	127	20.1
Toy	21	26.8
Cartoon	24	30.7
Manipulation	23	14.7
Total	**195**	**21.5**

Our video dataset contains four categories: Toy, Cartoon, Home, and Manipulation, with a total of 956 videos. Each video has a resolution of 1280×704 (720p) at 24 FPS and lasts approximately 5 seconds (121 frames). The Cartoon and Toy categories are non-realistic, synthetic scenarios designed to test models on stylized and imaginative content, while the Home and Manipulation categories include realistic scenes covering everyday objects and natural environments. By combining both realistic and non-realistic videos, our dataset provides a diverse benchmark for evaluating text-to-video retrieval models across various visual domains. Table 1 summarizes the number of main query descriptions for each video category in our GenState-AI dataset (after manual verification and filtering). The dataset contains four categories (Home, Toy, Cartoon and Manipulation), with a total of 195 main query descriptions. Notably, the Home category has the largest number of main queries, while Toy and Cartoon/Manipulation have fewer.

In our default setting, we generate 5-second clips at 720p using text-only conditioning, focusing on stable scenes (e.g., home, desk, kitchen counter). Since all clips are independently generated, GenState-AI should be understood as a prompt-controlled benchmark rather than a pixel-controlled counterfactual video benchmark.

GPT-4 generated the prompt triplets under a shared standardized instruction.

Prompts follow a simple, controlled structure designed to isolate semantic and temporal variations:

- Main video example: "On a shelf there are two apples. A hand adds one more."
- Semantic hard negative: "On a shelf there are two oranges. A hand adds one more."
- Temporal hard negative: "On a shelf there are two apples. A hand removes one."

Table 2: Statistics of our video dataset. Each video has a resolution of 1280×704 (720p) at 24 FPS and lasts about 5 seconds (121 frames).

Category	#Videos	Resolution	FPS	Duration (s)	Size
Home	368	1280×704	24	5	2.53 GB
Toy	221	1280×704	24	5	1.01 GB
Cartoon	150	1280×704	24	5	522 MB
Manipulation	217	1280×704	24	5	1.92 GB
Total	**956**				**5.98 GB**

This prompt structure is designed to encourage coherent motion while avoiding complex interactions such as pouring, cutting, or complex human actions. Clips inconsistent with the intended prompt or target state are removed through manual verification.

Table 2 summarizes the statistics of each category.

3.3 Example Prompts with AI-Generated Videos

Appendix A presents representative examples from each group of our benchmark. For every query prompt, the examples show the corresponding main video alongside its temporal and semantic hard negatives. These triplets illustrate the core difficulty of GenState-AI: despite being visually similar, the negatives differ from the main video either in object identity or in state evolution, requiring retrieval models to possess both semantic understanding and fine-grained temporal reasoning.

3.4 Manual Check and Correction

Although Wan2.2-TI2V-5B provides controllable video generation, it can occasionally produce clips that deviate from the intended prompt or fail to realize the target state transition. To ensure the reliability of GenState-AI, we conduct a structured human verification process. Five annotators (four M.S. students in AI and one Ph.D. researcher) divided the clips into five non-overlapping subsets, with exactly one reviewer per clip. They applied shared criteria for prompt consistency, required objects and actions, visibility of the intended transition, and the absence of severe visual artifacts; this protocol does not support inter-annotator agreement.

Main-clip verification (query validity). For each query description, annotators first inspect its *main* clip and verify whether the video faithfully expresses the intended state-change description. If the main clip is incorrect or ambiguous, we discard

only this main clip (and regenerate a new main candidate if needed), because an invalid main clip would invalidate the query–video alignment and make subsequent triplet-based evaluation uninterpretable.

Hard negatives as global distractors. In contrast, we do not enforce a strict *pairwise* constraint between a hard negative and a particular main clip during manual checking. Temporal and semantic hard-negative clips are retained even when the corresponding main clip is discarded, since these clips remain valid videos that can serve as *global negatives* (distractors) for other queries in the candidate pool. This design choice both increases the diversity of negatives and explains why the total number of verified clips is not necessarily a multiple of three.

Triplet integrity without asymmetric replacement. When forming query-specific triplets for triplet-ordering analyses, we retain only query groups for which the main, temporal, and semantic clips are all present in the evaluation export. We do not replace a missing component with a clip from another query, because that would introduce asymmetric examples and could bias retrieval evaluation. The resulting export contains 154 complete triplets; the remaining 41 main-query descriptions remain part of the full candidate-pool retrieval evaluation but are excluded from triplet-level metrics.

Across approximately 1,800 generated video clips, a total of 956 valid clips remain after manual verification. The total time to generate a single video on an RTX 5090 is approximately 7 minutes. These verified clips support both (i) constructing a core set of query-specific triplets for triplet-based evaluation and (ii) enriching the candidate pool with additional global negative videos for more challenging retrieval.

4 Evaluation Protocol

4.1 Standard Retrieval Setup

For each text query q, its mapped main clip v^{main} is the sole relevant target. Recall@K ranks it against all videos in the corresponding evaluation pool; other main clips and all available temporal or semantic hard negatives are ordinary non-relevant candidates.

We adopt standard text-to-video retrieval metrics. To emphasize state-awareness, we also report the proposed triplet metrics and analyze error types.

4.2 MLLM Baselines on GenState-AI

We evaluate representative MLLM-based retrieval baselines on GenState-AI from three complementary paradigms. First, we build a likelihood-scoring text-to-video baseline on top of **VideoChat-Flash-Qwen2-7B_res448** (denoted as **VCF-Lik**). For each candidate video v, we extract frame-level visual tokens with a frozen vision backbone and condition the MLLM on these tokens to score a text query q via teacher forcing. Concretely, we use the length-normalized token log-likelihood (equivalently, the negative cross-entropy) as the relevance score,

$$s_{\text{lik}}(q, v) = \frac{1}{|q|} \sum_{i=1}^{|q|} \log p_\theta(q_i \mid q_{<i}, v), \quad (1)$$

and rank candidates by $s_{\text{lik}}(q, v)$.

Second, we adopt a two-stage pipeline based on **Qwen3-VL**, where **Qwen3-VL-Embedding** performs first-stage retrieval and **Qwen3-VL-Reranker** conducts second-stage re-ranking (Li et al., 2026b). We evaluate scale-matched 2B and 8B embedding/reranker configurations; we denote them as **Qwen3-VL (2B)** and **Qwen3-VL (8B)**.

Finally, we include **LanguageBind** (Zhu et al., 2024), a contrastive multimodal representation model that aligns video and text in a shared embedding space through language-based semantic alignment. We use the fine-tuned video checkpoint *LanguageBind_Video_FT* and rank candidates by text–video embedding similarity. This baseline represents a single-stage dual-encoder retrieval paradigm without an explicit cross-candidate reranking step.

4.3 Triplet Hard-Negative Ordering Metrics

For a query whose mapped main, temporal-negative, and semantic-negative clips are all available, GenState-AI defines the query-specific *triplet*

$$\left(v^{\text{main}},\ v^{\text{temp}},\ v^{\text{sem}}\right),$$

where v^{main} is the ground-truth video, v^{temp} is a temporal hard negative, and v^{sem} is a semantic hard negative. Given any retrieval model that produces a scalar score $s(q, v)$ (e.g., a likelihood, embedding-similarity, or reranking score), we can study not only the global rank of v^{main}, but also its *relative ordering* with respect to the two hard negatives.

We distinguish two evaluation contexts. Recall@K is computed from the standard retrieval

Table 3: Performance comparison of the baselines across different subsets on GenState-AI. The Avg. row reports the macro average over the four subsets. Best results are highlighted in **bold**. Recall@K uses all videos in each corresponding subset pool and retains all 195 queries; the three triplet accuracies use the 154 complete triplets and exclude 41 incomplete groups.

Subset	Baseline	R@1	R@5	R@10	Acc_{temp}	Acc_{sem}	Acc_{tri}
Toy	VCF-Lik	0.381	0.714	0.857	0.619	0.810	0.524
	Qwen3-VL (2B)	0.571	**0.952**	0.952	0.667	**0.905**	0.667
	Qwen3-VL (8B)	**0.667**	**0.952**	**1.000**	**0.762**	**0.905**	**0.714**
	LanguageBind	0.571	0.905	0.952	0.619	0.857	0.619
Cartoon	VCF-Lik	0.250	0.708	0.708	0.500	0.833	0.417
	Qwen3-VL (2B)	0.708	**1.000**	**1.000**	0.792	0.875	0.750
	Qwen3-VL (8B)	0.667	**1.000**	**1.000**	0.792	0.958	0.750
	LanguageBind	**0.833**	0.958	**1.000**	**0.833**	**1.000**	**0.833**
Home	VCF-Lik	0.386	0.512	0.591	0.682	0.920	0.648
	Qwen3-VL (2B)	0.528	0.850	0.953	0.852	0.943	0.830
	Qwen3-VL (8B)	**0.685**	**0.929**	**0.984**	**0.909**	0.977	**0.898**
	LanguageBind	0.441	0.890	0.969	0.716	**0.989**	0.716
Manipulation	VCF-Lik	**0.478**	0.870	0.957	**0.952**	0.810	**0.810**
	Qwen3-VL (2B)	0.348	0.783	**0.957**	0.524	0.762	0.524
	Qwen3-VL (8B)	0.304	**0.957**	**0.957**	0.619	**0.857**	0.524
	LanguageBind	0.391	0.913	0.913	0.571	**0.857**	0.524
Avg.	VCF-Lik	0.374	0.701	0.778	0.688	0.843	0.599
	Qwen3-VL (2B)	0.539	0.896	0.965	0.709	0.871	0.693
	Qwen3-VL (8B)	**0.581**	**0.960**	**0.985**	**0.770**	0.924	**0.721**
	LanguageBind	0.559	0.916	0.958	0.685	**0.926**	0.673

ranking over all videos in the corresponding evaluated subset. In contrast, the following triplet metrics are computed as a diagnostic ordering task over only the query-specific triplet

$$\mathcal{T}_i = \{v_i^{\text{main}}, v_i^{\text{temp}}, v_i^{\text{sem}}\}.$$

That is, each model scores the three paired clips for the same query and the three clips are ranked relative to one another. For two-stage systems, this diagnostic evaluation bypasses the first-stage $K = 20$ candidate filter and directly scores the three triplet clips with the final scoring or reranking module. Therefore, the triplet metrics are not mathematically constrained by full-pool Recall@K.

Recall@K retains all 195 query descriptions. The triplet accuracies and relative-order histograms use only the 154 queries with complete main, temporal-negative, and semantic-negative mappings; 41 incomplete groups are excluded.

Let $\text{rank}_{\mathcal{T}}(q_i, v)$ denote the rank (1 is best) of candidate v among the three videos in $\mathcal{T}_i$. We define the following GenState-AI specific metrics:

Main-vs-Temporal accuracy. The probability that the model correctly prefers the main video over its temporal hard negative:

$$\text{Acc}_{\text{temp}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big[\text{rank}_{\mathcal{T}}(q_i, v_i^{\text{main}}) < \text{rank}_{\mathcal{T}}(q_i, v_i^{\text{temp}})\Big]. \tag{2}$$

In the reported triplet diagnostics, $N = 154$, corresponding to the queries with complete main, temporal-negative, and semantic-negative mappings.

Main-vs-Semantic accuracy. Analogously, the probability that the main video is ranked ahead of its semantic hard negative:

$$\text{Acc}_{\text{sem}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big[\text{rank}_{\mathcal{T}}(q_i, v_i^{\text{main}}) < \text{rank}_{\mathcal{T}}(q_i, v_i^{\text{sem}})\Big]. \tag{3}$$

Triplet consistency. A stricter metric that requires the main video to be ranked ahead of *both* hard negatives. For brevity, let $r_i^{\text{m}} = \text{rank}_{\mathcal{T}}(q_i, v_i^{\text{main}})$, $r_i^{\text{t}} = \text{rank}_{\mathcal{T}}(q_i, v_i^{\text{temp}})$, and $r_i^{\text{s}} = \text{rank}_{\mathcal{T}}(q_i, v_i^{\text{sem}})$. Then

$$\text{Acc}_{\text{tri}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[r_i^{\text{m}} < r_i^{\text{t}} \wedge r_i^{\text{m}} < r_i^{\text{s}}\big]. \tag{4}$$

Relative-order histogram. Beyond scalar accuracies, we also report the empirical distribution of three-way orders within each triplet. For example:

$$v^{\text{main}} \succ v^{\text{temp}} \succ v^{\text{sem}},$$
$$v^{\text{temp}} \succ v^{\text{main}} \succ v^{\text{sem}},$$
$$v^{\text{sem}} \succ v^{\text{main}} \succ v^{\text{temp}}, \text{ etc.}$$

where $a \succ b$ means $\text{rank}_{\mathcal{T}}(q, a) < \text{rank}_{\mathcal{T}}(q, b)$. This histogram reveals whether the model tends to systematically favor temporal or semantic hard negatives over the main video, and whether errors are dominated by temporal confusion (main ranked behind v^{temp}) or semantic confusion (main ranked behind v^{sem}).

These triplet-level metrics are specific to the GenState-AI design and complement standard Recall@K: they directly measure whether a model correctly distinguishes the intended main state transition from carefully controlled temporal and semantic alternatives.

5 Experiments

5.1 Implementation Details

We evaluate VCF-Lik, Qwen3-VL (2B/8B), and LanguageBind under a unified 5-frame full-video protocol. Qwen3-VL uses scale-matched Qwen3-VL-Embedding-2B/8B first-stage retrievers with corresponding 2B/8B second-stage rerankers; LanguageBind uses the same 5 uniformly sampled frames for comparability with Qwen3-VL and the last-frame ablation. For full-pool Qwen3-VL retrieval, the embedding model ranks all videos in the evaluated subset and passes the top-K candidates with $K = 20$ to the corresponding reranker; the final ranking reranks these 20 candidates and retains the remaining candidates in first-stage order. For Recall@K, the mapped main clip is the sole positive, and the candidate pool contains all videos in the corresponding evaluated subset for all three baselines; all 195 queries are retained. Acc_{temp}, Acc_{sem}, and Acc_{tri} are computed over the query-specific triplet as defined in Sec. 4.3. These metrics use the 154 complete triplets and exclude 41 incomplete groups. All Avg. rows are macro averages over the four subsets.

5.2 Overall Results on GenState-AI

Table 3 summarizes the performance of the four retrieval systems across the GenState-AI benchmark. All Recall results use the corresponding complete subset pool for every baseline. Appendix D reports an optional 956-video VCF-Lik stress test, and Appendix C reports LanguageBind frame-sampling sensitivity.

Scaling and Remaining Challenges. The results show that GenState-AI is sensitive to model capability while still exposing structured failures. Scaling the Qwen3-VL configuration from 2B to 8B, with scale-matched embedding and reranking models, improves the macro-average R@1 from 0.539 to 0.581, while Acc_{tri} increases only modestly from 0.693 to 0.721. This scale-matched comparison changes both the embedding and reranking components rather than isolating the reranker effect; higher full-pool retrieval scores do not automatically imply reliable triplet ordering. However, the improvement is not uniform across domains: on *Manipulation*, Qwen3-VL (8B) obtains a lower R@1 than Qwen3-VL (2B) (0.304 vs. 0.348), while both achieve the same triplet consistency (0.524). In contrast, VCF-Lik performs substantially better, reaching 0.478 R@1 and 0.810 Acc_{tri}. Thus, the larger scale-matched Qwen3-VL configuration attains higher macro-average retrieval scores, while fine-grained action-dependent temporal grounding remains challenging.

Performance Comparison. Table 3 shows that no single baseline dominates all evaluation axes. **Qwen3-VL (8B)** achieves the best macro-average retrieval performance (R@1: 0.581; R@5: 0.960; R@10: 0.985), suggesting that the larger scale-matched Qwen3-VL configuration is effective at selecting a single best candidate when the correct video survives first-stage retrieval. **LanguageBind**, however, obtains the best results on the *Cartoon* subset (R@1: 0.833; Acc_{tri}: 0.833), indicating that its language-aligned video embedding space is robust for stylized content. In contrast, **VCF-Lik** remains strongest on *Manipulation* in both R@1 (0.478) and Acc_{tri} (0.810), highlighting that likelihood-based matching can better preserve fine-grained action and end-state ordering when hand-object interactions are subtle and action-dependent. Overall, the systems exhibit complementary strengths across domains, indicating that GenState-AI captures diverse failure modes beyond static appearance matching.

Triplet Consistency. Across subsets, semantic discrimination is generally easier than temporal discrimination for all baselines (i.e., $\text{Acc}_{\text{sem}} >$

Acc_{temp} in most cases), aligning with the intuition that substituting objects or attributes often changes global appearance more than subtle end-state differences. This pattern is especially clear for **LanguageBind**: it achieves a high macro-average Acc_{sem} (0.926), but its Acc_{temp} remains much lower (0.685), with a pronounced gap on *Manipulation* (0.571 vs. 0.857). **Qwen3-VL (8B)** narrows this gap on average (Acc_{temp}: 0.770; Acc_{sem}: 0.924), but still suffers on *Manipulation*, where Acc_{tri} is 0.524. These results reveal a distinction between broad semantic matching and decisive temporal endpoint verification: stronger models improve retrieval, yet paired temporal negatives still expose subset-specific weaknesses. Overall, the results support the central motivation of GenState-AI: modern baselines can retrieve globally plausible videos, and the larger scale-matched Qwen3-VL configuration attains higher macro-average top-rank accuracy, but reliable state-aware grounding remains uneven across domains.

Value of Triplet Diagnostics. Recall@K and triplet consistency capture complementary aspects of performance. LanguageBind attains macro-average R@5/R@10 of 0.916/0.958 on all 195 queries, whereas its Acc_{temp} is 0.685 on the 154 complete triplets. These aggregate results concern different query sets, but illustrate that high global recall alone does not ensure correct temporal ordering. Despite lower macro-average Recall@K than Qwen3-VL, VCF-Lik achieves higher Manipulation Acc_{tri} (0.810 vs. 0.524 for both Qwen3-VL variants), showing what standard retrieval metrics alone would miss about state-sensitive ordering.

5.3 Diagnostic Ordering and Qualitative Trends

Beyond scalar triplet consistency, LanguageBind's relative-order distribution shows that its top-ranked errors are dominated by temporal hard negatives. Among the 154 complete triplets, the main video ranks first in 107 (69.5%) cases, a temporal negative in 45 (29.2%), and a semantic negative in 2 (1.3%). Temporal negatives rank first most often in Toy and Manipulation (38.1% in each). Appendix E provides the full ordering distributions and representative failures, illustrating the same challenge: a plausible clip can outrank the target despite conflicting with the queried motion or action.

5.4 Ablation Study: Does the End-State Alone Suffice?

Since our dataset emphasizes explicit final states, a natural question is whether models can solve it by inspecting only the terminal frame. We therefore compare the default full-video setting (5 uniformly sampled frames) with a *Last Frame Only* setting for Qwen3-VL and LanguageBind. This ablation does not by itself prove temporal sequence reasoning; instead, it tests whether terminal-frame appearance matching is sufficient for the evaluated models. Full subset-level results are provided in Appendix B.

At the macro-average level, full-video input improves R@1 and Acc_{tri} for all evaluated systems, although subset-level metrics do not improve uniformly (Appendix B). For Qwen3-VL (8B), R@1/Acc_{tri} fall from 0.581/0.721 to 0.399/0.515 under last-frame-only input; LanguageBind falls from 0.559/0.673 to 0.435/0.574. On Toy, Qwen3-VL (8B) drops from 0.667 to 0.238 in R@1 and from 0.762 to 0.238 in Acc_{temp}. These results show that terminal-frame appearance alone is insufficient for this benchmark.

6 Conclusion

We introduced GenState-AI, a prompt-controlled benchmark of AI-generated videos and verified state-change queries across four visual domains. Its evaluation separates standard retrieval over complete subset-level video pools from diagnostic ordering over query groups with complete main, temporal-negative, and semantic-negative mappings. This distinction reveals that a model may retrieve the correct video near the top globally while still preferring a paired clip with an incorrect state transition. The scale-matched Qwen3-VL (8B) configuration improves macro-average retrieval and triplet consistency over 2B, but temporal/state confusions remain. Qwen3-VL (8B) achieves the strongest overall retrieval and triplet consistency, whereas VCF-Lik performs notably better on Manipulation triplets, demonstrating complementary strengths across scoring paradigms. GenState-AI is therefore intended as a focused diagnostic benchmark rather than an exhaustive measure of video understanding. Its verified candidate pools, controlled hard negatives, and explicit evaluation protocol provide a reproducible basis for developing retrieval systems that better distinguish semantic relevance from state-transition correctness.

Limitations

GenState-AI is prompt-controlled rather than pixel-controlled. Although hard-negative prompts are minimally edited and all clips are manually verified, independently generated videos may still vary in low-level appearance, lighting, pose, camera framing, or generation artifacts. All clips use Wan2.2-TI2V-5B, so residual generator-specific cues may remain and cross-generator generalization remains untested. The current benchmark focuses on four common state-centric subsets and 195 verified queries; it is intended as a diagnostic testbed rather than a comprehensive coverage of all event types, domains, or video generation models. Finally, our experiments cover representative open-source retrieval paradigms, including likelihood scoring, embedding–reranking, and dual-encoder matching, but do not exhaustively evaluate all closed-source or larger-scale multimodal systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62376178).

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Dian Chen and Philipp Krähenbühl. 2022. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17222–17231.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. 2023. Videollm: Modeling video sequence with large language models. *Preprint*, arXiv:2305.13292.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, Xuhui Jiang, and Jian Guo. 2025. Vlm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8140–8158.

Yang Du, Yuqi Liu, and Qin Jin. 2024. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5260–5269.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. 2024. Ltx-video: Realtime video latent diffusion. *Preprint*, arXiv:2501.00103.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, Xianfang Zeng, Xinhao Zhang, Gang Yu, Yuhe Yin, Qiling Wu, Wen Sun, Kang An, Xin Han, Deshan Sun, and 35 others. 2025. Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model. *Preprint*, arXiv:2503.11251.

Chan Hur, Jeong-hun Hong, Dong-hun Lee, Dabin Kang, Semin Myeong, Sang-hyo Park, and Hyeyoung Park. 2025. Narrating the video: Boosting text-video retrieval via comprehensive utilization of frame-level captions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24077–24086.

Dohwan Ko, Ji Soo Lee, Minhyuk Choi, Zihang Meng, and Hyunwoo J Kim. 2025. Bidirectional likelihood estimation with multi-modal large language models for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22263–22273.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Minghan Li and Eric Gaussier. 2022. Bert-based dense intra-ranking and contextualized late interaction via multi-task learning for long document retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2347–2352.

Minghan Li and Eric Gaussier. 2024. Domain adaptation for dense retrieval and conversational dense retrieval through self-supervision by meticulous pseudo-relevance labeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5247–5259.

Minghan Li, Eric Gaussier, Juntao Li, and Guodong Zhou. 2024. Evirerank: Adaptive evidence construction for long-document llm reranking. *arXiv preprint arXiv:2411.06254*.

Minghan Li, Diana Nicoleta Popa, Johan Chagnon, Yagmur Gizem Cinar, and Eric Gaussier. 2023. The power of selecting key blocks with local pre-ranking for long document information retrieval. *ACM Transactions on Information Systems*, 41(3):1–35.

Minghan Li, Yishuai Zhang, Tianrui Lv, Siqi Zhao, Miyang Luo, Ercong Nie, and Guodong Zhou. 2026a. A survey of long-document retrieval in the plm and llm era. *ACM Transactions on Information Systems*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026b. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *Preprint*, arXiv:2601.04720.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2026c. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pages 109089–109117.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pages 44215–44234.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. 2024a. Sora: A review on background, technology, limitations, and opportunities of large vision models. *Preprint*, arXiv:2402.17177.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024b. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772.

Ziming Liu, Di Luo, Yilun Xu, Tommi Jaakkola, and Max Tegmark. 2023. Genphys: From physical processes to generative models. *arXiv preprint arXiv:2304.02637*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2024. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M de Melo, Benjamin Van Durme, and Rama Chellappa. 2025. Video-colbert: Contextualized late interaction for text-to-video retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19691–19701.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2023. Make-a-Video: Text-to-Video Generation Without Text-Video Data. In *International Conference on Learning Representations*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, and 43 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2024. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16551–16560.

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

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. 2024. Carebench: A fine-grained benchmark for video captioning and retrieval. *arXiv preprint arXiv:2501.00513*.

Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, Yihang Duan, Xinyu Lyu, and Heng Tao Shen. 2025. Text-video retrieval with global-local semantic consistent learning. *IEEE Transactions on Image Processing*.

Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. 2025. Controlvideo: conditional control for one-shot text-driven video editing and beyond. *Science China Information Sciences*, 68(3):132107.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.

Bin Zhu, Bin Lin, Munan Ning, YANG YAN, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Zhang, Zhifeng Li, Wei Liu, and Yuan Li. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, volume 2024, pages 9588–9608.

A Qualitative Dataset Examples

Figure 2 provides one representative triplet from each subset. These examples are intended to make the construction protocol visually inspectable: the temporal negative targets a conflicting state-transition outcome, while the semantic negative changes the queried object, attribute, or action target.

Because the clips are generated independently from edited prompts, they should be read as prompt-controlled rather than pixel-controlled counterfactuals. The examples show that the paired videos preserve the broad scene and task setting while still exhibiting natural generation-level variation in pose, lighting, object appearance, and composition.

B Full Last-Frame Ablation Results

Table 4 reports the complete subset-level results for the last-frame ablation. All Avg. rows are macro averages over the four subsets. Recall@K uses each corresponding complete subset pool and all 195 query descriptions, while the triplet accuracies use the 154 complete query-specific triplets.

The full table expands the summary reported in the main experiments. Across models, replacing the 5-frame video input with only the final frame generally reduces both retrieval recall and triplet consistency, especially on state-centric subsets such as Toy and Home. This supports the conservative conclusion that terminal-frame appearance alone is insufficient for reliably solving GenState-AI, although it does not by itself prove that models perform full temporal-order reasoning.

C LanguageBind Frame-Sampling Sensitivity

Table 5 compares LanguageBind with the unified 5-frame protocol used in the main experiments against its 8-frame default setting. We treat this as a frame-sampling sensitivity analysis rather than a separate main baseline.

The 5-frame setting is used in the main tables so that LanguageBind is evaluated under the same visual evidence budget as VCF-Lik and Qwen3-VL. The 8-frame setting is retained here because it reflects the model's default usage and helps estimate whether the main conclusion depends on this implementation choice. The difference is modest at the macro-average level, but subset-level changes are visible: 8 frames improve Home while reducing Cartoon and Manipulation, suggesting that frame sampling can interact with scene type and generated-video style.

D VCF-Lik Dataset-Wide Retrieval Stress Test

Table 6 reports an optional 956-video cross-subset stress test that is not used in Table 3. Only Cartoon decreases because non-Cartoon candidates displace correct clips across the Recall cutoffs; Toy, Home, and Manipulation are unchanged. The subset-pool columns rank each query against videos in its corresponding subset, whereas the 956-video columns use the full dataset as the candidate pool.

E Diagnostic Ordering and Failure Visualizations

Figure 3 summarizes the relative-order distributions of the 154 complete triplets under LanguageBind. The desired ordering places the main video ahead of both negatives; deviations reveal whether the model is more often confused by temporal/state conflicts or semantic distractors. Across all triplets, the main video is ranked first in 107/154 (69.5%)

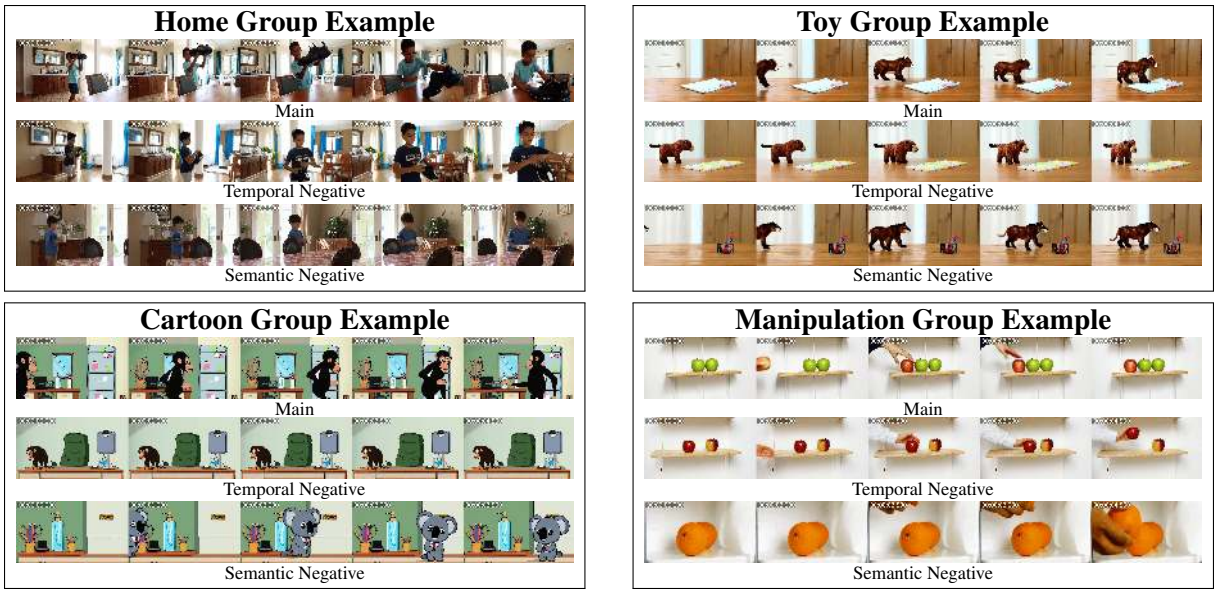


Figure 2: Qualitative examples of generated videos and paired hard negatives across all four subsets. Each image is a horizontally concatenated video strip; each group shows the main video, temporal hard negative, and semantic hard negative.

Table 4: Full subset-level ablation comparing full-video context (5 frames) against static last-frame context. Recall@K uses all videos in each corresponding subset pool and retains all 195 queries, whereas the triplet accuracies use the 154 complete query-specific triplets; 41 incomplete query groups are excluded.

Subset	Model	Context	R@1	R@5	R@10	Acc_{temp}	Acc_{sem}	Acc_{tri}
Cartoon	Qwen3-VL (2B)	Full	0.708	1.000	1.000	0.792	0.875	0.750
	Qwen3-VL (2B)	Last	0.625	0.833	0.917	0.708	0.917	0.667
	Qwen3-VL (8B)	Full	0.667	1.000	1.000	0.792	0.958	0.750
	Qwen3-VL (8B)	Last	0.625	0.958	1.000	0.667	0.958	0.667
	LanguageBind	Full	0.833	0.958	1.000	0.833	1.000	0.833
	LanguageBind	Last	0.750	0.958	0.958	0.750	0.917	0.750
Home	Qwen3-VL (2B)	Full	0.528	0.850	0.953	0.852	0.943	0.830
	Qwen3-VL (2B)	Last	0.409	0.780	0.921	0.739	0.898	0.705
	Qwen3-VL (8B)	Full	0.685	0.929	0.984	0.909	0.977	0.898
	Qwen3-VL (8B)	Last	0.472	0.835	0.953	0.761	0.898	0.727
	LanguageBind	Full	0.441	0.890	0.969	0.716	0.989	0.716
	LanguageBind	Last	0.299	0.717	0.858	0.614	0.830	0.545
Manip.	Qwen3-VL (2B)	Full	0.348	0.783	0.957	0.524	0.762	0.524
	Qwen3-VL (2B)	Last	0.174	0.609	0.957	0.524	0.857	0.476
	Qwen3-VL (8B)	Full	0.304	0.957	0.957	0.619	0.857	0.524
	Qwen3-VL (8B)	Last	0.261	0.696	0.870	0.476	0.762	0.429
	LanguageBind	Full	0.391	0.913	0.913	0.571	0.857	0.524
	LanguageBind	Last	0.261	0.826	0.957	0.571	0.905	0.476
Toy	Qwen3-VL (2B)	Full	0.571	0.952	0.952	0.667	0.905	0.667
	Qwen3-VL (2B)	Last	0.476	0.952	0.952	0.524	0.810	0.524
	Qwen3-VL (8B)	Full	0.667	0.952	1.000	0.762	0.905	0.714
	Qwen3-VL (8B)	Last	0.238	0.857	0.905	0.238	0.810	0.238
	LanguageBind	Full	0.571	0.905	0.952	0.619	0.857	0.619
	LanguageBind	Last	0.429	0.810	0.857	0.524	0.810	0.524
Avg.	Qwen3-VL (2B)	Full	0.539	0.896	0.965	0.709	0.871	0.693
	Qwen3-VL (2B)	Last	0.421	0.793	0.937	0.624	0.870	0.593
	Qwen3-VL (8B)	Full	0.581	0.960	0.985	0.770	0.924	0.721
	Qwen3-VL (8B)	Last	0.399	0.836	0.932	0.536	0.857	0.515
	LanguageBind	Full	0.559	0.916	0.958	0.685	0.926	0.673
	LanguageBind	Last	0.435	0.828	0.908	0.615	0.865	0.574

Table 5: LanguageBind frame-sampling sensitivity. The 5-frame setting is used in the main tables for protocol consistency. Recall@K uses each corresponding complete subset pool; triplet accuracies use the 154 complete triplets.

Subset	Frames	R@1	R@5	R@10	Acc_{temp}	Acc_{sem}	Acc_{tri}
Toy	5	0.571	0.905	0.952	0.619	0.857	0.619
Toy	8	0.571	0.905	0.952	0.619	0.857	0.619
Cartoon	5	0.833	0.958	1.000	0.833	1.000	0.833
Cartoon	8	0.750	1.000	1.000	0.792	1.000	0.792
Home	5	0.441	0.890	0.969	0.716	0.989	0.716
Home	8	0.504	0.921	0.984	0.750	0.989	0.750
Manip.	5	0.391	0.913	0.913	0.571	0.857	0.524
Manip.	8	0.304	0.870	0.957	0.476	0.905	0.429
Avg.	5	0.559	0.916	0.958	0.685	0.926	0.673
Avg.	8	0.532	0.924	0.973	0.659	0.938	0.647

Table 6: VCF-Lik Recall using the Table 3 subset pools versus all 956 videos.

	Subset pool			956-video pool		
Subset	R@1	R@5	R@10	R@1	R@5	R@10
Toy	0.381	0.714	0.857	0.381	0.714	0.857
Cartoon	0.250	0.708	0.708	0.208	0.417	0.625
Home	0.386	0.512	0.591	0.386	0.512	0.591
Manip.	0.478	0.870	0.957	0.478	0.870	0.957

cases, a temporal negative in 45/154 (29.2%), and a semantic negative in 2/154 (1.3%).

Under LanguageBind, temporal negatives rank first in 38.1% of complete triplets in both Toy and Manipulation. Figure 4 shows two VCF-Lik errors: a same-object, opposite-motion Toy negative outranks the target, as does an action-mismatched Home negative in a similar scene.

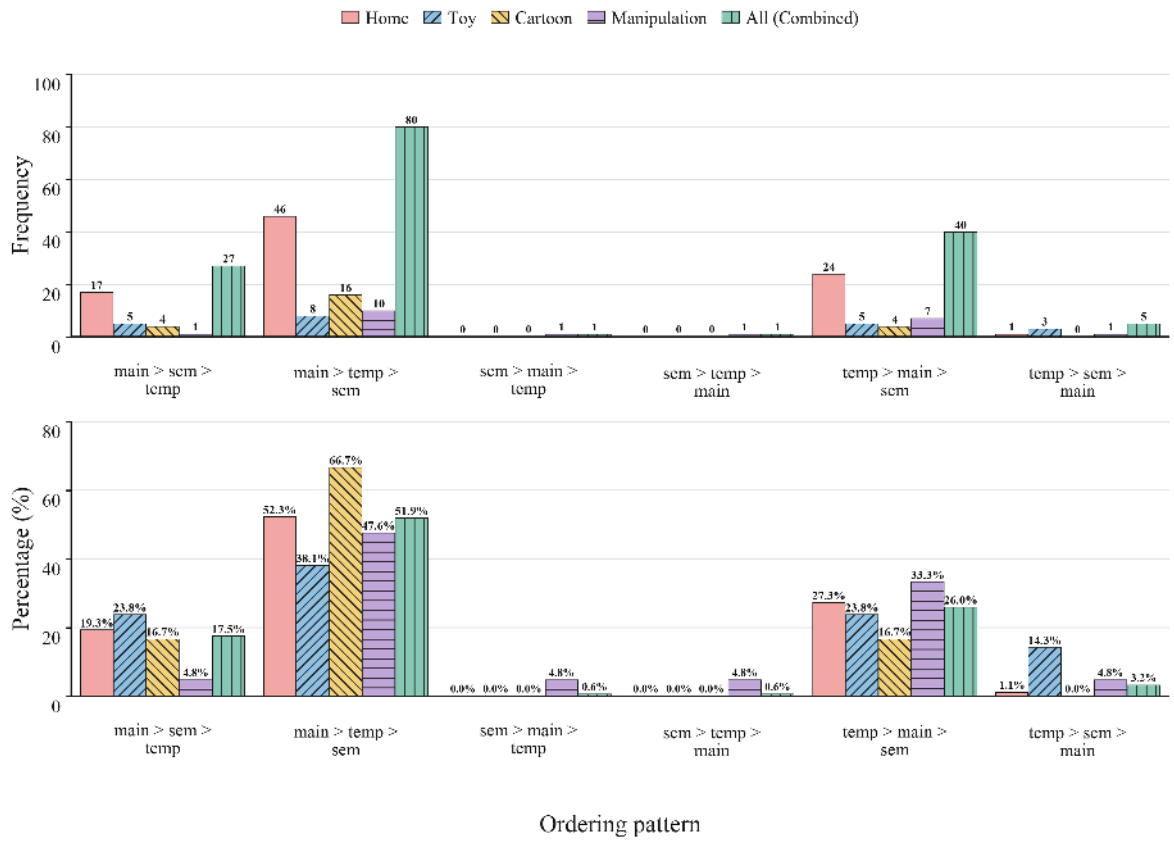


Figure 3: Comparison of relative-order distributions across subsets under **LanguageBind**. Counts and percentages are computed from the complete triplets available for each subset (154 in total).



Figure 4: Representative triplet-ordering failures under **VCF-Lik**; higher-ranked negatives appear above the main video. In **Toy**, Rank 1 is a temporal negative showing the same rubber duck moving from left to right beside a gift box, while the Rank 2 target shows the duck moving from right to left; Rank 3 is a semantic negative featuring a jumping cartoon owl. Thus, the same-object clip with opposite motion outranks the target. In **Home**, Rank 1 is a semantic negative showing a girl moving a potted plant to a windowsill, Rank 2 is a temporal negative showing her lifting and laying down a blue blanket, and the Rank 3 target shows her picking up a book. This ordering illustrates difficulty distinguishing the target action from hard negatives in a similar visual context.