The Shape of Time: Video-Token Contrast for Temporal Understanding in VideoLMs

Anonymous ACL submission

Abstract

Seeing frames in order does not mean representing time. Modern VideoLMs receive ordered video streams, yet their main supervision acts on generated text rather than video-token representations where event dynamics should first emerge. This mismatch allows models to learn temporal answers from shortcuts such as objects, scenes, and language priors, without requiring internal video representations to capture event progression. To address this, we propose VT-Contrast, a representation-level temporal counterfactual objective for VideoLMs. Its design asks where temporal supervision should act and what temporal differences it should expose. VT-Contrast supervises selected late-layer last-frame video tokens, where temporal information is expected to be integrated before language generation, and contrasts order-preserving views with same-video reordered counterfactuals graded by Kendall tau distance. It requires no architectural changes, is compatible with diverse VideoLM training tasks, and improves overall performance across temporal understanding benchmarks. Our code is available at *https://anonymous.4open.science/r/VT-Contrast*.

1 Introduction

Video Language Models (VideoLMs) (Lin et al., 2024; Maaz et al., 2024) are increasingly expected to reason about how events unfold over time (Li et al., 2026; Qin et al., 2025), rather than merely recognize visual content. For example, two clips may contain nearly the same objects and scenes, while expressing opposite events under different temporal progressions: opening versus closing, entering versus leaving, or assembling versus disassembling. Resolving such order-dependent distinctions requires models to capture the progression of visual states over time.

However, standard VideoLM training does not directly enforce this requirement. While VideoLMs receive temporal information through ordered visual tokens, supervision is usually applied only to the generated text (Zhang et al., 2023, 2024c). The language modeling loss specifies the desired answer, yet does not require the resulting visual representations to reflect the event progression that supports it. For order-sensitive videos, this is a weak constraint: clips with similar visual evidence but different temporal order may still lead to correct responses based on static cues or answer priors. As a result, response-level supervision may leave video representations underconstrained along the dimension most relevant to event progression, even when the final responses are correct.

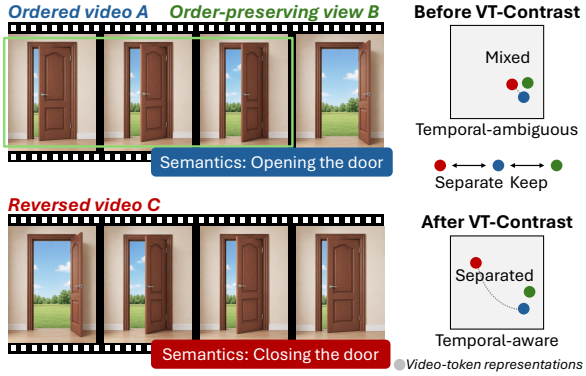

Figure 1: Motivation of VT-Contrast. Temporal order changes video semantics; VT-Contrast separates temporally corrupted negatives while keeping order-consistent positives close to the anchor in the video-token space.

A more direct form of temporal supervision is therefore needed before response generation. Such supervision should make temporal order explicit in the internal video-token representation space, rather than relying solely on the final textual response. Prior approaches move in related directions but do not directly impose this constraint. Temporal-aware architectures (Hu et al., 2024) improve temporal aggregation in visual or intermediate modules, but do not directly supervise temporal-

order representations after video tokens enter the language-model space. Ordering-based auxiliary tasks (Wu et al., 2025) expose models to shuffled segments, but still rely on task-level responses and can be less compatible with existing training objectives. Reconstruction objectives (Tong et al., 2022) offer more direct visual supervision for representations, but often require extra branches and costly pixel-level targets.

A representation-level temporal signal should distinguish videos primarily by their temporal order rather than by their visual content. We instantiate this idea by constructing multiple views from the same video: order-preserving views keep the original event progression, whereas reordered views retain the same visual evidence but change how the event unfolds. As illustrated in Figure 1, such order changes can alter event semantics, yet existing VideoLMs may still map the corresponding video-token representations close together. We formulate these reordered views as temporal counterfactuals and propose **VT-Contrast**, a lightweight objective that separates them from order-preserving views in the video-token space. This makes the learning signal depend on temporal order rather than video identity or static appearance cues.

Beyond this counterfactual construction, VT-Contrast must decide where the supervision should act and how strong the temporal signal should be. For supervision placement, we apply the objective to selected late-layer last-frame video tokens, which are expected to aggregate information from preceding frames before language generation. This choice is motivated by prior findings that temporal information emerges progressively inside VideoLM representations, rather than being uniformly available at all layers (Zhang et al., 2025b; Shi et al., 2026). For temporal signal strength, we avoid treating all reorderings as equivalent. We use Kendall tau distance (Kendall, 1938) to quantify pairwise order violations, enabling VT-Contrast to construct graded temporal counterfactuals that yield more effective contrastive supervision.

VT-Contrast is jointly optimized with the standard language modeling objective, requiring no architectural modification. By directly shaping video-token geometry, it complements response-level supervision: the model is still trained to generate correct answers, while its internal video representations are encouraged to reflect event progression. Experiments on multiple temporal understanding benchmarks show consistent overall gains, especially on order-sensitive tasks. Our contributions are summarized as follows:

- We introduce a representation-level temporal supervision perspective for VideoLMs, arguing that temporal order should be reflected in video-token representations rather than only supervised through textual responses.

- We propose VT-Contrast, a lightweight temporal counterfactual objective for VideoLM training that uses same-video temporally reordered sequences as negatives. It applies targeted supervision to selected-layer last-frame token representations and controls negative difficulty with Kendall tau distance.

- Evaluations on multiple temporal benchmarks show consistent overall improvements. VT-Contrast integrates into standard VideoLM training without changing the original task format, enabling flexible temporal enhancement across diverse video understanding scenarios.

2 Related Works

2.1 Video Language Models

Recent VideoLMs (Zhang et al., 2024b; Lin et al., 2024) typically extend LLM-based multimodal understanding from images to videos by encoding frames or clips into visual tokens, projecting them into LLM feature space, and generating textual responses autoregressively (Liu et al., 2023; Maaz et al., 2024). Representative systems include both closed-source models (Comanici et al., 2025; Achiam et al., 2023) and open-source models (Bai et al., 2025; Chen et al., 2024; Li et al., 2024). Despite differences in model scale, data, and training recipes, their training signal is primarily imposed on generated text tokens rather than on intermediate video-token representations (Zhang et al., 2023, 2024c; Feng et al., 2026). Although language loss can update visual components, it does not directly constrain video-token representations to distinguish temporal order changes.

2.2 Temporal Understanding

Temporal understanding is fundamental to VideoLMs, as videos require models to reason over event order, motion direction, state changes, and other relations beyond static visual recognition (Zhang et al., 2024a; Chandrasegaran et al.,

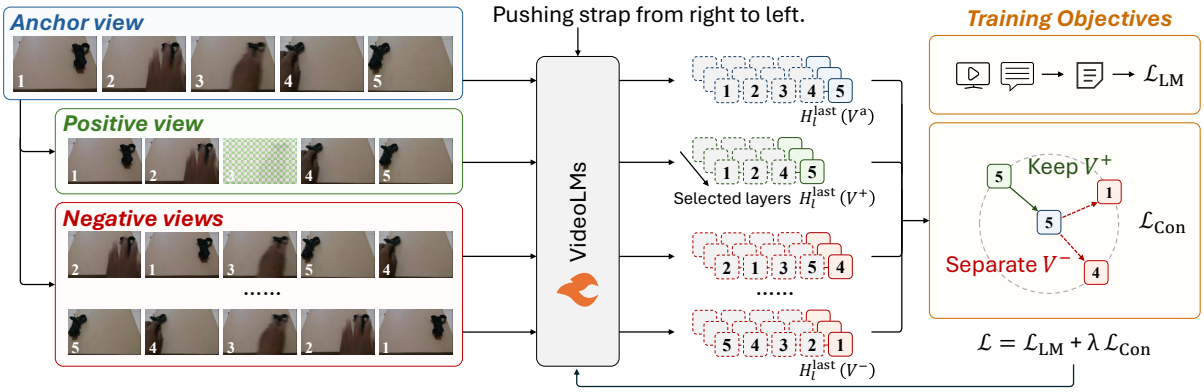

Figure 2: Overview of VT-Contrast. For each input video, we construct an anchor view, an order-preserving positive view, and temporal counterfactual negative views. A shared VideoLM processes these views to obtain selected video-token representations for contrastive supervision. VT-Contrast keeps temporally consistent views close and separates negative views, and is jointly optimized with the standard language modeling loss.

2024). Recent benchmarks show that current VideoLMs still struggle with fine-grained temporal semantics, especially when videos share similar visual content but differ in temporal order or event progression (Liu et al., 2024; Shangguan et al., 2025). To improve temporal understanding, prior work has explored temporal-aware architectures, temporal reasoning transfer, and auxiliary ordering tasks (Hu et al., 2024; Fateh et al., 2024). For instance, T3 (Li et al., 2025) transfers temporal reasoning from synthetic text tasks, while Visual Jigsaw (Wu et al., 2025) trains models to recover shuffled visual order. While these efforts highlight the importance of temporal modeling, directly shaping video-token representations for temporal sensitivity remains less explored.

2.3 Contrastive Learning

Contrastive learning is widely used for representation learning by pulling positive pairs together and pushing negative samples apart in feature space (He et al., 2020; Chen et al., 2020). In vision-language learning, CLIP-style objectives align visual and textual representations (Radford et al., 2021), and recent studies further extend contrastive training to VLM-based multimodal embedding models (Jiang et al., 2024; Meng et al., 2025). Beyond image-text alignment, contrastive objectives have also been applied to video representation learning, including self-supervised methods that learn general spatiotemporal representations from video clips and temporal perturbations (Pan et al., 2021; Qian et al., 2021; Dave et al., 2022; Dorkenwald et al., 2022). However, these methods mainly focus on self-supervised pretraining of standalone video encoders, rather than representation-level temporal supervision for VideoLMs.

3 Method

We propose VT-Contrast, a lightweight temporal counterfactual learning method that enhances temporal order sensitivity in VideoLMs through video-token supervision. It provides direct representation-level supervision by contrasting temporally consistent views with reordered views from the same video. We then describe how VT-Contrast is integrated into VideoLM training.

3.1 Problem Setup

Given a common video-language training dataset, each sample consists of a video $V = \{v_1, \ldots, v_T\}$, a textual query Q, and a target response $Y = \{y_1, \ldots, y_N\}$, where v_t denotes the t-th sampled frame. A VideoLM processes V to obtain video-token representations, which are used as visual context for the language model. The response is predicted autoregressively as:

$$p_\theta(Y \mid V, Q) = \prod_{i=1}^{N} p_\theta(y_i \mid y_{<i}, V, Q), \quad (1)$$

where p_θ denotes the token distribution predicted by the VideoLM with parameters θ.

Standard VideoLM training minimizes the negative log-likelihood of the target response:

$$\mathcal{L}_{\text{LM}} = -\sum_{i=1}^{N} \log p_\theta(y_i \mid y_{<i}, V, Q). \quad (2)$$

Although this response-level objective aligns videos with textual answers, it supervises video-token representations only indirectly through the

output loss. Consequently, these representations are not explicitly encouraged to distinguish fine-grained temporal order changes, which can allow the model to rely on static visual cues or language priors when they are sufficient for response generation. We therefore introduce an auxiliary contrastive objective based on temporal counterfactual views to directly supervise video-token representations during VideoLM training.

3.2 Temporal Counterfactual Learning

As illustrated in Figure 2, VT-Contrast constructs temporally consistent and order-disrupted views from the same input video. The order-disrupted views serve as temporal counterfactuals: they preserve the visual content of the original video while altering its event progression. Instead of relying on textual responses for temporal supervision, our method contrasts these video views directly in the representation space. This makes temporal order differences explicit in video-token representations without changing the response format.

Contrastive view construction. We instantiate this idea by constructing three views from each training video $V = \{v_1, \ldots, v_T\}$: an anchor view V^a, an order-consistent positive view V^+, and an order-disrupted negative view V^-. The anchor defines the reference temporal progression, the positive preserves this progression under a mild sampling perturbation, and the counterfactual negative view changes the progression while keeping the sampled visual content fixed. Together, these views allow the contrastive objective to compare temporal consistency against order disruption under controlled visual content.

▷ **Anchor:** The anchor view V^a is the reference chronological view. It is obtained by uniformly sampling K frames from the original video while preserving their temporal order. Formally, let $i_1 < i_2 < \cdots < i_K$ denote the sampled frame indices; the anchor view is defined as:

$$V^a = \{v_{i_1}, v_{i_2}, \ldots, v_{i_K}\}. \tag{3}$$

▷ **Positive:** The positive view V^+ is an order-consistent counterpart to the anchor view. It preserves the temporal progression and main visual content of V^a while introducing a mild sampling variation. In this work, we instantiate this perturbation by randomly dropping one frame from V^a. Let $d \in \{1, \ldots, K\}$ denote the dropped position; the positive view is defined as:

$$V^+ = \{v_{i_1}, \ldots, v_{i_{d-1}}, v_{i_{d+1}}, \ldots, v_{i_K}\}. \tag{4}$$

Since the remaining frames preserve the original chronological order, V^a and V^+ are treated as a positive pair.

▷ **Negative:** The negative view V^- is designed to isolate temporal order as the main contrasting factor. It reuses the same sampled frames as the anchor view, so that objects, scenes, and local visual patterns are largely preserved while the chronological order is disrupted. This same-video construction reduces shortcuts based on video identity or static appearance, making temporal order the primary difference between V^a and V^-.

Specifically, we apply a non-identity temporal permutation π to the anchor sequence:

$$V^- = \{v_{i_{\pi(1)}}, v_{i_{\pi(2)}}, \ldots, v_{i_{\pi(K)}}\}, \quad \pi \neq \mathrm{id}. \tag{5}$$

Thus, V^- serves as a temporal counterfactual negative that differs from V^a in frame order.

Contrastive objective. To realize temporal counterfactual learning, we train the anchor representation to stay close to its order-consistent positive view and away from same-video counterfactual negatives. In this way, the contrastive objective turns temporal order changes into explicit constraints on video-token representations. Specifically, we adopt an InfoNCE objective (Oord et al., 2018). For each anchor view, we construct M counterfactual negative views $\{V_m^-\}_{m=1}^M$. Let $r(U)$ denote the video-token representation extracted from any constructed view U. We compute the positive and negative similarity scores as:

$$\begin{aligned} s^+ &= \mathrm{sim}\big(r(V^a), r(V^+)\big), \\ s_m^- &= \mathrm{sim}\big(r(V^a), r(V_m^-)\big). \end{aligned} \tag{6}$$

The contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{con}} = -\log \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \sum_{m=1}^M \exp(s_m^-/\tau)}, \tag{7}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature parameter.

3.3 Emergence-aware Temporal Supervision

The temporal counterfactual objective should not be applied blindly to all video tokens or all reordered views. We therefore design the supervision

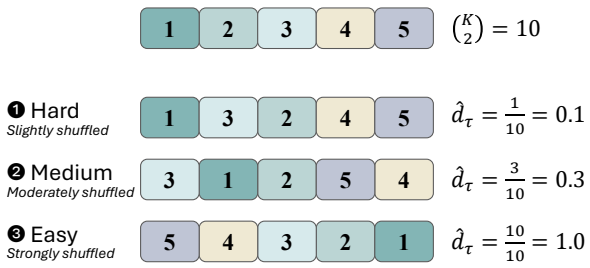

Figure 3: Kendall-tau-based grading of temporal counterfactual negatives. Smaller $\hat{d}_\tau$ yields harder counterfactuals closer to the original order, while larger $\hat{d}_\tau$ indicates stronger order disruption and easier contrasts.

to be both targeted and graded, so that the representation constraint focuses on informative temporal features while keeping the counterfactual comparisons at an appropriate difficulty.

Targeted video-token supervision. We now describe how the contrastive representation $r(U)$ is extracted for each constructed view. Each view is processed by the shared VideoLM to obtain layer-wise video-token features. Motivated by the observation that different layers capture different stages of multimodal information flow and last-frame features can aggregate temporal cues across the video (Zhang et al., 2025b; Shi et al., 2026), we use last-frame representations from selected layers $\mathcal{S}$ for contrastive supervision. This targeted design avoids dense constraints over all layers and video tokens while focusing on compact, temporally informative features.

Specifically, for each view U of the original video V and selected layer $l \in \mathcal{S}$, we take the last-frame video-token features, denoted as $H_l^{\text{last}}(U)$. We first apply mean pooling over the last-frame tokens at each selected layer and then average the resulting features across selected layers:

$$r(U) = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \text{MeanPool}\left(H_l^{\text{last}}(U)\right), \quad (8)$$

where $\text{MeanPool}(\cdot)$ averages the last-frame video tokens into a single feature vector. Since different VideoLM backbones vary in depth and representation structure, the selected layer set $\mathcal{S}$ is model-dependent. We specify the exact layers in the implementation details.

Graded temporal counterfactuals. Temporal counterfactuals should differ from the anchor in temporal order, but the degree of order violation affects the learning signal. Smaller violations create harder fine-grained contrasts, while larger violations produce more explicit but easier counterfactuals. Therefore, we use Kendall tau distance to grade the strength of order violation, allowing us to control the signal and construct effective temporal contrasts, as illustrated in Figure 3.

Given a permutation π_m over the K frames of the anchor view, the Kendall tau distance counts the number of pairwise order inversions:

$$d_\tau(\pi_m) = \sum_{1 \le i < j \le K} \mathbb{I}[\pi_m(i) > \pi_m(j)], \quad (9)$$

where $\mathbb{I}[\cdot]$ is the indicator function. Since the maximum number of inversions is $K(K-1)/2$, we normalize the distance as:

$$\hat{d}_\tau(\pi_m) = \frac{2 d_\tau(\pi_m)}{K(K-1)}. \quad (10)$$

The normalized distance $\hat{d}_\tau(\pi_m) \in [0, 1]$ measures how strongly the counterfactual violates the original temporal order. Smaller values indicate that the permutation remains close to the original order and is harder to distinguish, whereas larger values indicate stronger order disruption.

During training, we sample each counterfactual permutation π_m from a controlled range:

$$\delta_{\min} < \hat{d}_\tau(\pi_m) \le \delta_{\max}, \quad (11)$$

where $\delta_{\min}$ and $\delta_{\max}$ define the allowed degree of order violation. If the number of valid permutations in the range is insufficient, we repeatedly sample with replacement until obtaining the required M negatives. In our implementation, we use $0 < \hat{d}_\tau(\pi_m) \le 0.5$. This favors counterfactuals that introduce order violations while remaining relatively close to the original sequence, making them harder to distinguish.

3.4 Optimization Objective

We keep the standard language modeling objective: it provides semantic supervision for the original temporally ordered video-query pair and preserves the model's generation ability. VT-Contrast is added as an auxiliary objective, providing direct temporal supervision on video-token representations. The final training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{con}}, \quad (12)$$

where λ controls the balance between language modeling and temporal contrastive supervision. In all experiments, we set $\lambda = 0.1$, which enables stable joint optimization.

Frames	Model	TOMATO	TempCompass				Vinoground		
			Y/N	MCQ	Caption	Match	Text	Video	Group
8	LLaVA-OneVision1.5-4B	21.09	62.86	56.71	50.00	68.26	20.60	23.80	3.60
	InternVL3.5-4B	28.71	72.08	68.73	60.93	79.44	48.20	28.20	16.00
	Visual Jigsaw	26.01	73.50	71.77	61.93	79.51	44.80	25.00	13.20
	Qwen3.5-0.8B	26.01	63.64	58.80	48.35	70.39	26.60	22.00	7.00
	Qwen3.5-2B	25.74	71.02	67.34	56.64	77.45	34.00	20.40	7.60
	Qwen3.5-4B	33.42	77.62	71.90	49.50	83.77	50.00	31.80	19.60
	Ours-0.8B	**27.29** +1.28	**64.98** +1.34	**59.49** +0.69	**50.05** +1.70	70.26 -0.13	**29.40** +2.80	**22.20** +0.20	**7.20** +0.20
	Ours-2B	**31.47** +5.73	**74.97** +3.95	**70.00** +2.66	**64.17** +7.53	**80.17** +2.72	**41.20** +7.20	**25.00** +4.60	**11.40** +3.80
	Ours-4B	**36.46** +3.04	**79.33** +1.71	**74.94** +3.04	**50.80** +1.30	**85.30** +1.53	**57.20** +7.20	29.00 -2.80	**20.00** +0.40
16	LLaVA-OneVision1.5-4B	20.55	62.25	57.47	51.15	69.33	20.00	23.80	4.80
	InternVL3.5-4B	30.12	72.89	70.70	61.68	80.44	53.60	30.80	16.80
	Visual Jigsaw	25.13	74.60	72.78	62.72	80.64	47.20	28.00	15.40
	Qwen3.5-0.8B	25.81	64.04	59.94	49.75	70.73	27.60	21.20	5.80
	Qwen3.5-2B	26.55	72.48	69.18	58.83	78.84	37.40	23.00	9.60
	Qwen3.5-4B	35.58	77.66	74.68	50.90	84.56	57.80	33.40	22.40
	Ours-0.8B	**27.63** +1.82	**65.27** +1.23	**60.95** +1.01	**50.15** +0.40	**71.39** +0.66	**31.00** +3.40	**22.60** +1.40	**8.80** +3.00
	Ours-2B	**31.20** +4.65	**76.07** +3.59	**71.46** +2.28	**65.77** +6.94	**82.24** +3.40	**43.40** +6.00	**24.80** +1.80	**12.80** +3.20
	Ours-4B	**37.67** +2.09	**79.49** +1.83	**76.20** +1.52	50.55 -0.35	**85.63** +1.07	**59.80** +2.00	31.60 -1.80	21.00 -1.40
32	LLaVA-OneVision1.5-4B	20.49	62.49	56.71	49.85	68.80	19.80	23.40	4.80
	InternVL3.5-4B	29.65	73.46	70.76	62.03	80.71	52.80	32.60	19.40
	Visual Jigsaw	26.48	74.52	72.78	61.93	80.24	48.80	30.80	18.60
	Qwen3.5-0.8B	26.08	64.57	60.89	51.50	71.99	28.00	20.00	5.80
	Qwen3.5-2B	27.02	72.97	69.43	58.13	79.57	38.20	22.60	9.20
	Qwen3.5-4B	36.39	78.60	74.62	48.65	84.96	59.20	33.80	21.80
	Ours-0.8B	**28.17** +2.09	**66.08** +1.51	**62.22** +1.33	50.10 -1.40	**72.12** +0.13	**30.60** +2.60	**25.60** +5.60	**8.80** +3.00
	Ours-2B	**30.93** +3.91	**75.95** +2.98	**71.71** +2.28	**65.52** +7.39	**82.44** +2.87	**44.60** +6.40	**25.40** +2.80	**13.20** +4.00
	Ours-4B	**38.41** +2.02	**79.54** +0.94	**76.58** +1.96	**50.75** +2.10	**86.09** +1.13	**60.40** +1.20	**34.00** +0.20	**22.60** +0.80

Table 1: Main results on TOMATO, TempCompass, and Vinoground under different frame settings. Our 4B model performs better than recent 4B MLLMs on most benchmarks. Our method also improves over the corresponding Qwen3.5 models in most settings, showing that VT-Contrast strengthens temporal understanding without architectural changes. Small colored numbers denote absolute gains over the corresponding Qwen3.5 model.

4 Experiments

In this section, we evaluate the proposed method through extensive experiments. We first describe the experimental setup, present the main results, and then analyze key design choices and learned representations. We also provide qualitative analysis in Appendix A.5.

4.1 Experimental Settings

Baselines. We compare models trained with our method against four representative baselines. InternVL3.5 (Wang et al., 2025) and LLaVA-OneVision1.5 (An et al., 2025) are recent open-source multimodal large language models; we report their 4B variants. Visual Jigsaw (Wu et al., 2025) is a 7B model tailored for temporal understanding and trained for order prediction with GRPO (Shao et al., 2024). Finally, we include Qwen3.5 (Qwen Team, 2026), the base model used in our training, and evaluate its 0.8B, 2B, and 4B variants to study performance across scales.

Datasets. For training, we use the Something-Something V2 (SSv2) (Goyal et al., 2017) training split, a large-scale dataset for fine-grained human-object action understanding.

For evaluation, we use three benchmarks focused on temporal capability: TOMATO (Shangguan et al., 2025), TempCompass (Liu et al., 2024), and Vinoground (Zhang et al., 2024a). TOMATO evaluates temporal reasoning over event progression and state changes, while TempCompass covers action, speed, direction, and event order. Vinoground focuses on fine-grained video-caption alignment under subtle temporal variations. We report accuracy across all benchmarks, where higher values indicate better temporal understanding.

Implementation details. Our models are initialized from Qwen3.5 (Qwen Team, 2026), a recent state-of-the-art open-source model, at three scales: 0.8B, 2B, and 4B. We fine-tune them on the SSv2 (Goyal et al., 2017) training split under a video question answering formulation, using

Setting	λ	TOMATO	TempCompass				Vinoground		
			Y/N	MCQ	Caption	Match	Text	Video	Group
LM only	0	29.44	72.85	68.86	62.72	78.84	36.40	22.00	9.20
+ VT-Contrast	0.1	**31.47**	**74.97**	**70.00**	**64.17**	**80.17**	**41.20**	**25.00**	**11.40**
Gain	–	*+2.03*	*+2.12*	*+1.14*	*+1.45*	*+1.33*	*+4.80*	*+3.00*	*+2.20*

Table 2: Effect of the temporal counterfactual objective. Compared with LM-only training, VT-Contrast consistently improves across benchmarks, showing the benefit of direct temporal supervision on video-token representations. Gains denote absolute improvements over the LM-only baseline.

Token Representation	TOMATO	TempCompass				Vinoground		
		Y/N	MCQ	Caption	Match	Text	Video	Group
Last query token	**31.81**	71.50	68.86	59.23	77.45	33.40	**25.80**	9.80
All frame tokens	30.26	73.99	**70.19**	62.92	79.77	40.20	21.40	9.80
Last-frame tokens	31.47	**74.97**	70.00	**64.17**	**80.17**	**41.20**	25.00	**11.40**

Table 3: Effect of token representation for temporal supervision. Last-frame video tokens achieve the best overall performance, suggesting that targeted supervision on temporally aggregated video-token representations is more effective for VT-Contrast. Bold and underlined numbers indicate the best and second-best results, respectively.

"What action is happening in the video?" as the prompt and the action label as the target response. We jointly optimize the language modeling loss and VT-Contrast, setting $\lambda = 0.1$. For VT-Contrast, we sample 16 negatives per instance with Kendall tau distance in $(0, 0.5]$, uniformly sample 8 frames per video, and use last-frame video tokens. We apply VT-Contrast to layers 21–24 out of 24 layers for the 0.8B and 2B models, and to layers 25–28 out of 32 layers for the 4B model. All models are trained for 2,250 steps using the same global batch size of 16 and a learning rate of 1×10^{-5}.

During evaluation, we use the lmms-eval framework (Zhang et al., 2025a), disable thinking mode, and adopt greedy decoding for deterministic results. To reduce formatting-related evaluation errors, we use prompts with standardized answer formats; details are provided in Appendix B.

4.2 Main Results

Table 1 reports the main results under different frame budgets. Following lmms-eval, frames are first sampled at the default FPS and capped by the specified frame budget, so the reported frame setting is an upper bound and some videos may use fewer frames. Compared with recent open-source VideoLMs, our 4B model achieves stronger results on most temporal benchmarks, with a particularly clear advantage on TOMATO.

Compared with the corresponding Qwen3.5 base models, our method brings consistent gains across most model scales and frame settings. The gains are particularly notable for the 2B model: under the 8-frame setting, it improves TOMATO by 5.73% and TempCompass Y/N by 3.95%, showing clear benefits on order-sensitive tasks. The 0.8B and 4B models also show clear improvements, suggesting that the proposed objective remains effective across model scales, with further room for scale-specific tuning. At the same time, the TempCompass Caption results suggest that some subtasks remain partly dependent on the base model capability under greedy decoding; in our results, the 2B model performs better than the 4B model on Caption and also obtains a larger gain.

4.3 Ablation Studies

We ablate the key designs of VT-Contrast, including the temporal counterfactual objective, targeted video-token representations, and counterfactual difficulty control. All experiments use the 2B model under the 8-frame setting. Additional analyses of layer selection, the number of negatives, and temperature are provided in Appendices A.2, A.3, and A.4, respectively.

Temporal counterfactual objective. Table 2 ablates the temporal counterfactual objective. Compared with LM-only training, VT-Contrast improves performance across the three benchmarks, with gains on every metric. These consistent gains suggest that the objective provides useful supervision beyond standard language modeling, with larger improvements on TOMATO and Vinoground Text. We use $\lambda = 0.1$ by default, with additional loss-weight experiments in Appendix A.1.

Counterfactual Difficulty	TOMATO	TempCompass				Vinoground		
		Y/N	MCQ	Caption	Match	Text	Video	Group
Full negatives	30.86	73.62	69.94	63.22	79.71	39.20	**26.60**	10.80
Simple negatives	30.80	73.75	68.92	63.82	80.04	39.20	24.20	11.00
Hard negatives	**31.47**	**74.97**	**70.00**	**64.17**	**80.17**	**41.20**	25.00	**11.40**

Table 4: Effect of temporal counterfactual difficulty. We compare full $(0, 1]$, simple $(0.5, 1]$, and hard $(0, 0.5]$ negatives based on Kendall-tau distance. The best performance from hard counterfactuals shows that Kendall-tau-based grading helps select more useful temporal counterfactuals.

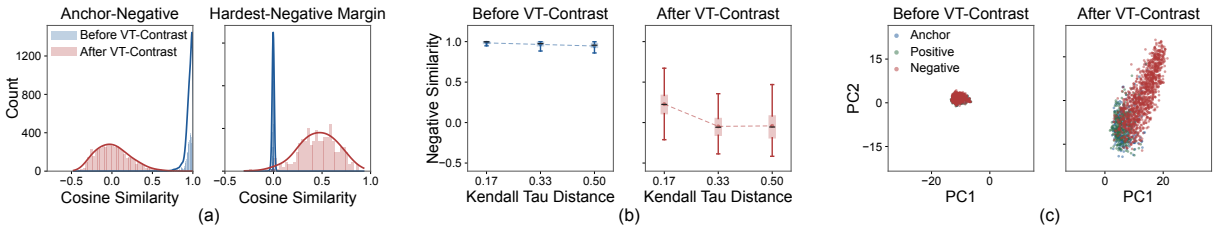

Figure 4: Representation analysis of VT-Contrast. (a) Similarity distributions: VT-Contrast separates hardest negatives from positives while keeping anchor-positive similarities relatively high. (b) Distance-similarity comparison: negative similarity decreases more clearly with Kendall-tau distance after training. (c) PCA visualization: anchor, positive, and negative representations become more distinguishable after training.

Targeted video-token selection. Table 3 studies where to apply VT-Contrast. Compared with supervising the last query token used for answer generation, supervising video-token representations achieves substantially better results on TempCompass and Vinoground Group. This highlights the importance of applying temporal counterfactual supervision directly in the video-token space. Among video-token choices, last-frame tokens outperform full-frame tokens overall, suggesting that the final-frame representations provide a compact target that better aggregates temporal information.

Temporal counterfactual difficulty. Table 4 ablates Kendall-tau-based negative selection. Hard negatives achieve the best overall performance across most benchmarks. Their advantage over full negatives suggests that a focused hard subset is better than mixing widely varying difficulties, while their advantage over simple negatives shows the benefit of more challenging perturbations. Together, these results demonstrate the importance of Kendall-tau-based grading for selecting effective temporal counterfactuals.

4.4 Representation Analysis

To understand how VT-Contrast improves temporal understanding, we analyze video-token representations on 300 SSv2 validation samples. As shown in Figure 4(a), before training, original videos and their reordered variants have very high cosine similarity and are mapped close in representation space. This supports our motivation that standard VideoLM representations are insufficiently sensitive to temporal order perturbations.

After training, reordered variants are less concentrated around the original representation, and the hardest-negative margin increases. Figure 4(b) relates anchor-negative similarity to Kendall-tau distance, showing a clearer decrease with stronger temporal perturbations. Figure 4(c) shows PCA, where negative representations become more distinguishable from anchor and positive views. Overall, these results show that same-video temporal counterfactuals provide effective supervision for order-sensitive video-token representations.

5 Conclusion

We propose VT-Contrast, a representation-level temporal supervision method for VideoLMs. The central idea is to shift temporal learning from response space to video-token space by contrasting order-preserving views with same-video reordered counterfactuals. Applied to targeted last-frame representations, VT-Contrast uses Kendall tau distance to grade the temporal disruption of counterfactuals, making temporal order explicit in internal representations. Experiments across temporal benchmarks show consistent overall gains, suggesting that shaping video-token representations is promising for improving temporal reasoning in VideoLMs.

Limitations

Although VT-Contrast shows consistent improvements, our validation is limited by computational resources to moderate-scale supervised fine-tuning on existing VideoLMs. Its effectiveness on larger and more diverse video-text data, such as longer videos, multi-event scenarios, and broader instruction-style tasks, is worth further exploration. Since VT-Contrast provides auxiliary supervision by reshaping existing video-token representations, its gains may still depend on the temporal modeling and generation ability of the base model. This is reflected by cases where larger models do not always perform better, such as TempCompass Caption, suggesting that the gains from VT-Contrast are affected by the base model's task-specific capability. These observations point to a promising direction of incorporating temporal counterfactual supervision into larger-scale pretraining to further strengthen VideoLMs' temporal capability.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. 2022. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406.

Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. 2022. Scvrl: Shuffled contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4132–4141.

Fawad Javed Fateh, Umer Ahmed, Hamza Khan, M Zeeshan Zia, and Quoc-Huy Tran. 2024. Video llms for temporal reasoning in long videos. *arXiv e-prints*, pages arXiv–2412.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2026. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, and 1 others. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Zi-Yuan Hu, Yiwu Zhong, Shijia Huang, Michael Lyu, and Liwei Wang. 2024. Enhancing temporal modeling of video llms via time gating. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2845–2856.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. 2025. Temporal reasoning transfer from text to video. In *International Conference on Learning Representations*, volume 2025, pages 74070–74102.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. 2026. Universal video temporal grounding with generative multi-modal large language models. *Advances in Neural Information Processing Systems*, 38:64426–64455.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and 1 others. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11205–11214.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974.

Hangyu Qin, Junbin Xiao, and Angela Yao. 2025. Question-answering dense video events. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 884–894.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2025. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations*, volume 2025, pages 7593–7734.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yumeng Shi, Quanyu Long, Yin Wu, and Wenya Wang. 2026. Causality matters: How temporal information emerges in video language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9006–9014.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *Advances in neural information processing systems*, 35:10078–10093.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. 2025. Visual jigsaw post-training improves mllms. *arXiv preprint arXiv:2509.25190*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. 2024a. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2025a. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024b. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025b. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19781–19791.

A Additional Analysis

In this appendix, we provide additional analysis of the proposed VT-Contrast. We study the effects of the contrastive weight λ, targeted layer selection, the number M of negative views, and contrastive temperature T. We also include qualitative examples to better illustrate the behavior of our model. Overall, these results show that VT-Contrast is relatively robust to common hyperparameter choices. Unless otherwise specified, we follow the main experimental setting and report results based on the 2B model with 8 input frames.

A.1 Contrastive Loss Weight

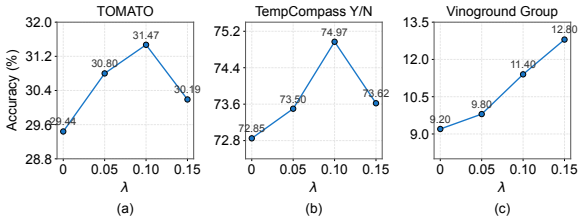

Figure 5: Effect of the contrastive loss weight λ. Adding VT-Contrast generally improves performance over $\lambda = 0$, while larger weights do not always help. We use $\lambda = 0.1$ as the default setting, which gives the best overall performance in this ablation.

We first study the effect of the contrastive loss weight λ. For TempCompass and Vinoground, we report TempCompass Y/N and Vinoground Group because the former is highly order-sensitive and the latter is the most representative Vinoground metric. As shown in Figure 5, adding VT-Contrast improves performance over $\lambda = 0$ on all three benchmarks. As λ increases from 0, the stronger contrastive signal generally improves temporal discrimination. However, after a certain point, a larger weight can hurt some tasks, likely because the contrastive objective starts to interfere with the model's general downstream performance. Therefore, we use $\lambda = 0.1$ as the default setting, which gives the best overall trade-off in these results.

A.2 Targeted Layer Selection

We further study contrastive layer selection using the same three representative tasks as Appendix A.1. As shown in Figure 6, different layer ranges lead to clearly different performance, indicating that where VT-Contrast is applied matters. A general trend is that middle-to-late layers perform better than early layers. This suggests that temporal information may be sufficiently aggregated in higher layers, while early-layer representations are still forming local visual features and are less suitable for our temporal counterfactual supervision. These results highlight the importance of targeted layer selection for obtaining strong performance. For the 2B model, we therefore apply VT-Contrast to layers 21–24. Although the best layer range may vary across model scales and deserves further exploration, later layers generally provide a strong and reliable default choice.

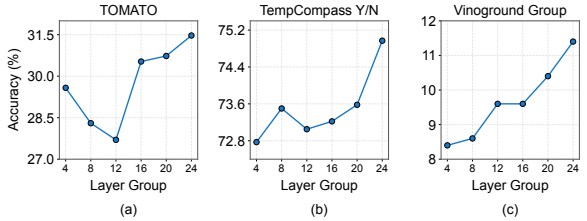

Figure 6: Effect of targeted layer selection. Different layer ranges show different effects, highlighting the importance of targeted layer selection. For the 2B model, we use layers 21–24, which perform best across the representative temporal tasks.

A.3 Number of Counterfactual Negatives

Table 5 studies the number of temporal counterfactual negatives. The number of negatives controls how many temporally perturbed same-video views are compared with each anchor in the InfoNCE objective. With only 4 negatives, the contrastive signal is limited, leading to weaker overall performance. Using 8 negatives already gives competitive results on several tasks, suggesting that moderate negative diversity is helpful. Further increasing the number to 16 provides richer temporal counterfactual comparisons and leads to the most stable results. Therefore, we use 16 counterfactual negatives as the default setting.

A.4 Contrastive Temperature

Table 6 studies the sensitivity to the contrastive temperature T. In the InfoNCE objective, T controls the sharpness of the similarity distribution: smaller values make the objective more sensitive to similarity differences, while larger values produce smoother contrastive signals. The results show that VT-Contrast is relatively robust to this hyperparameter, with different temperatures giving comparable performance across most tasks. Among them, $T = 0.10$ provides a strong trade-off, performing best on several tasks while remaining competitive on others. A smaller temperature sharpens the con-

Negative Number	TOMATO	TempCompass				Vinoground		
		Y/N	MCQ	Caption	Match	Text	Video	Group
4 samples	30.91	73.79	69.56	62.18	79.37	38.40	23.00	10.80
8 samples	**32.21**	73.42	69.43	**64.32**	79.97	39.00	23.80	10.20
16 samples	31.47	**74.97**	**70.00**	64.17	**80.17**	**41.20**	**25.00**	**11.40**

Table 5: Effect of the number M of counterfactual negatives. Using 4 negatives provides a relatively weak temporal counterfactual signal, while using 8 negatives already gives competitive results on several tasks. Increasing to 16 negatives further improves overall stability and achieves the best overall performance.

Temperature	TOMATO	TempCompass				Vinoground		
		Y/N	MCQ	Caption	Match	Text	Video	Group
0.06	29.99	73.58	**70.51**	**64.62**	79.24	**43.60**	23.40	11.20
0.08	31.27	73.54	69.68	63.67	**80.17**	39.20	24.60	**11.60**
0.10	**31.47**	**74.97**	70.00	64.17	**80.17**	41.20	**25.00**	11.40
0.12	30.39	74.56	69.68	62.57	79.91	40.60	23.40	11.40
0.14	29.65	73.05	70.13	63.72	79.31	40.60	22.60	10.80

Table 6: Sensitivity to the contrastive temperature T. Different temperatures yield comparable results across most tasks, indicating that VT-Contrast is not highly sensitive to this hyperparameter. We use $T = 0.10$ by default, which provides a strong overall trade-off.

trastive distribution, which can help some tasks but also leads to less stable results. By contrast, a larger temperature smooths the distribution and weakens the separation signal, leading to weaker overall performance. We therefore use $T = 0.10$ as the default setting.

A.5 Qualitative Analysis

Figure 7 presents qualitative examples from TempCompass Y/N, where each case contains an original video and its reversed counterpart. We choose this setting because the original and reversed videos share nearly the same visual content, while the questions are designed to test whether the model can distinguish changes in event direction and temporal order. As shown in the examples, Qwen3.5 is less sensitive to the temporal change between the original and reversed videos. It fails on fine-grained temporal transitions, including the child's turning direction, paper deformation, and the apple's visual state change, while our method adapts more reliably to the reversed temporal semantics. These cases suggest that VT-Contrast improves sensitivity to temporal progression and reduces reliance on static appearance cues.

B Prompt Details

For reproducibility, we list the prompts used for benchmark evaluation. Across all reported models and settings, we use the same prompt format within each benchmark to reduce prompt-induced variation. For each task, we present a prompt example illustrating the input format.

B.1 TOMATO

TOMATO is evaluated as a multiple-choice task. Following the default prompt in the lmms-eval framework, we further constrain the model to answer with a single option letter. This reduces overly open-ended responses and mitigates instability introduced by ChatGPT-based answer grading. An example prompt is shown below.

You will be provided with n separate frames uniformly sampled from a video, the frames are provided in chronological order of the video. Analyze these frames and provide the answer to the question about the video content. Answer the multiple-choice question about the video content.

You must use these frames to answer the multiple-choice question; do not rely on any externel knowledge or commonsense.

<question>
In which direction(s) did the person's hand move?
</question>

<options>
'A': 'Not moving at all', 'B': 'Left.', 'C': 'Right.', 'D': 'First to the right then to the left.', 'E': 'First to the left then to the right.'
</options>

Even if the information in these separate frames is not enough to answer the question, PLEASE TRY YOUR BEST TO GUESS AN ANSWER WHICH YOU THINK

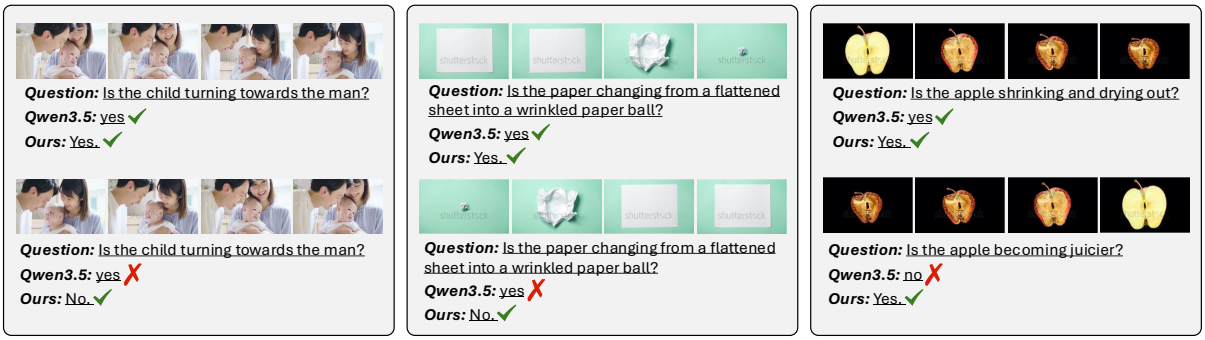

Figure 7: Qualitative examples on TempCompass Y/N with reversed video pairs. Each example contains an original video and its reversed counterpart, which share similar visual content but require different answers. Compared with Qwen3.5, our method better adapts its predictions to the changed temporal order.

WOULD BE THE MOST POSSIBLE ONE BASED ON THE QUESTION.

DO NOT GENERATE ANSWER SUCH AS 'NOT POSSIBLE TO DETERMINE.'

Your output must be exactly one uppercase option letter from A, B, C, D, E.

Answer:

B.2 TempCompass

For TempCompass, we use task-specific prompts for Yes or No, Multi Choice, Captioning, and Caption Matching. As with TOMATO, we explicitly specify the required answer format when applicable to ensure consistent evaluation across models. For Captioning, we follow the lmms-eval default and use gpt-3.5-turbo for GPT-based grading. Example prompts are provided below.

***Yes or No**:*
Is the man dunking?
Please answer yes or no:

***Multi Choice**:*
What is the man doing in the video?
A. dunking a basketball
B. dribbling a basketball
C. passing a basketball
Please only give the letter of the best option from A, B, C:

***Captioning**:*
You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly.
Information A: 'subject': 'man', 'action': 'dribbling a basketball'
Information B: 'subject': 'man', 'action': 'dunking a basketball'
Information C: 'subject': 'man', 'action': 'passing a basketball'
Generated Caption:

***Caption Matching**:*
Which description is a more suitable match for the video?
Option 1: The man is dribbling a basketball.
Option 2: A man is dunking a basketball.
Please only give the label of the best option from Option 1, Option 2:

B.3 Vinoground

For Vinoground, we use separate prompts for text-to-video and video-to-text matching. Depending on the matching direction, the model chooses between two candidate videos or captions and outputs only the corresponding option letter. Example prompts are shown below.

Text:
Which caption best describes this video?
A. a toddler picks up a water bottle and drinks before he plays around the grass field
B. a toddler plays around the grass field before he picks up a water bottle and drinks
Answer with the option's letter from the given choices directly. Please only output one English character.

***Video**:*
Which video segment matches this caption? Note: The video contains two segments separated by a 2-second black frame. Caption: a toddler plays around the grass field before he picks up a water bottle and drinks
A. First segment (before black frame)
B. Second segment (after black frame)
Answer with the option's letter from the given choices directly. Please only output one English character.