VideoPASTA: 7K Preference Pairs That Matter for Video-LLM Alignment

Yogesh Kulkarni **Pooyan Fazli**
Arizona State University
https://people-robots.github.io/VideoPASTA/

Abstract

Video-language models (Video-LLMs) excel at understanding video content but struggle with spatial relationships, temporal ordering, and cross-frame continuity. To address these limitations, we introduce **VideoPASTA** (**P**reference **A**lignment with **S**patio-**T**emporal-Cross Frame **A**dversaries), a framework that enhances Video-LLMs through targeted preference optimization. VideoPASTA trains models to distinguish accurate video representations from carefully crafted adversarial examples that deliberately violate spatial, temporal, or cross-frame relationships. With only 7,020 preference pairs and Direct Preference Optimization, VideoPASTA enables models to learn robust representations that capture fine-grained spatial details and long-range temporal dynamics. Experiments demonstrate that VideoPASTA is model agnostic and significantly improves performance, for example, achieving gains of up to +3.8 percentage points on LongVideoBench, +4.1 on VideoMME, and +4.0 on MVBench, when applied to various state-of-the-art Video-LLMs. These results demonstrate that targeted alignment, rather than massive pretraining or architectural modifications, effectively addresses core video-language challenges. Notably, VideoPASTA achieves these improvements without any human annotation or captioning, relying solely on 32-frame sampling. This efficiency makes our approach a scalable plug-and-play solution that seamlessly integrates with existing models while preserving their original capabilities.

1 Introduction

Recent advances in video language models (Video-LLMs) have enabled efficient video understanding and reasoning, achieving strong performance on tasks like captioning and question answering (Li et al., 2025a; Wang et al., 2022, 2024d; Bai et al., 2025; Chen et al., 2024c). However, these models typically rely on large, high-quality annotated

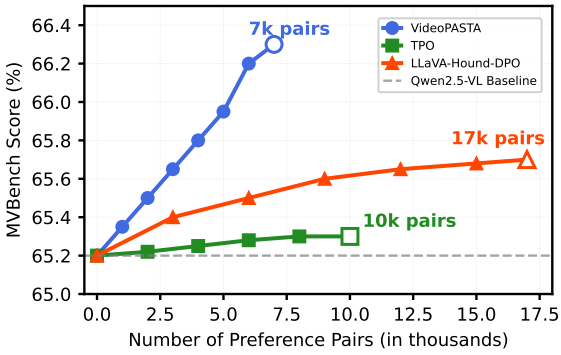


Figure 1: **With just 7k preference pairs**, VideoPASTA outperforms the Qwen2.5-VL (Bai et al., 2025) baseline, LLaVA-Hound (Zhang et al., 2025), and TPO (Li et al., 2025b) on MVBench, showing that targeted alignment surpasses models trained on larger datasets.

datasets and significant computing resources for training. Instruction tuning has emerged as a way to reduce data requirements by fine-tuning models on curated instruction-response pairs (Liu et al., 2023; Zhang et al., 2024d; Lin et al., 2024; Chen et al., 2024c; Wang et al., 2024b). While large instruction datasets have been generated using models like GPT-4V (Zhang et al., 2024d), improvements from training on these datasets remain limited. Video-LLMs still struggle with spatial misalignment, temporal incoherence, and cross-frame disconnections (Choong et al., 2024; Huang et al., 2025; Leng et al., 2024; Ma et al., 2024; Gunjal et al., 2024). Addressing these issues through human annotation is expensive, as it requires identifying examples with proper grounding and coherence. This suggests that merely scaling models and data alone is insufficient. Instead, the core challenge lies in achieving faithful alignment between model outputs and video content. Recent work has applied Direct Preference Optimization (DPO) to improve video-language alignment (Ahn et al., 2024a; Zhang et al., 2025; Li et al., 2025b; Rafailov et al., 2023). However, these methods often reinforce existing strengths through collecting more prefer-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32343–32368
November 4-9, 2025

ence data instead of targeting core weaknesses in Video-LLMs. They also rely on proprietary models (Zhang et al., 2025) or video captioning (Li et al., 2025b), limiting scalability.

This raises a key question: *How can we align Video-LLMs to understand spatial, temporal, and cross-frame relationships without human annotations, captions, or proprietary models, while remaining computationally efficient?* To address this, we introduce **VideoPASTA** (**P**reference **A**lignment with **S**patio-**T**emporal-Cross Frame **A**dversaries), a model-agnostic framework that improves video-language alignment using targeted preference pairs. VideoPASTA contrasts aligned ("preferred") responses with adversarial ones that capture three common failure modes in video understanding: (1) **spatial misalignment**, where responses misrepresent object relationships and interactions, (2) **temporal incoherence**, where responses violate the natural progression of events, and (3) **cross-frame disconnection**, where responses violate object persistence, character consistency, and narrative progression across more distant parts of a video. By combining DPO with structured preference data, VideoPASTA directly tackles these limitations in Video-LLMs. In summary, our contributions are as follows:

1. We introduce VideoPASTA, a novel, model-agnostic DPO framework that improves video-language alignment by addressing spatial misalignment, temporal incoherence, and cross-frame disconnection, without human annotations, captions, or proprietary models.
2. VideoPASTA sets a new efficiency benchmark, achieving strong results using just 7,020 preference pairs, far fewer than prior instruction tuning (1.3M) or preference datasets (17k).
3. Extensive evaluations on seven benchmarks show consistent, model-agnostic gains, with improvements of up to +3.8 percentage points on LongVideoBench, +4.1 on VideoMME, and +4.0 on MVBench.

2 Related Work

Video-LLMs. Despite advances in Video-LLMs (Li et al., 2024a; Wang et al., 2024b; Chen et al., 2024c), evaluations (Fu et al., 2025; Zhou et al., 2025; Liu et al., 2024b) reveal persistent challenges in three key areas. First, temporal reasoning, especially in long videos, remains difficult. Approaches like longer context (Shen et al., 2024; Zhang et al., 2024b), compression (Li et al., 2024c; Wang et al., 2024c), and training-free methods (Yang et al., 2025; Huang et al., 2025) improve token efficiency but not core understanding, while specialized methods (Chen et al., 2024b; Ren et al., 2024) demand heavy computation. Second, spatial misalignment leads to poor object localization and occlusion handling (Ranasinghe et al., 2024; Chen et al., 2024a). Third, cross-frame disconnection disrupts continuity and narrative coherence (Tan et al., 2024; Huang et al., 2024). Most methods address only one issue or rely on large-scale instruction tuning (Wang et al., 2024b; Zhang et al., 2024d; Lin et al., 2024), which fails to solve these core alignment problems. VideoPASTA uses DPO-based training on structured preference pairs to jointly address temporal, spatial, and cross-frame failures. By challenging models across all three dimensions, it achieves more comprehensive video-language alignment than conventional instruction tuning.

Video-Language Alignment. While reward modeling (Sun et al., 2024; Ahn et al., 2024b; Wang et al., 2024a) and self-training methods (Deng et al., 2024; Zohar et al., 2025; Kulkarni and Fazli, 2025b,a) aim to improve video-language alignment and reduce the need for manual annotations, existing approaches still face major limitations. Prior DPO applications, such as LLaVA-Hound-DPO (Zhang et al., 2025) and i-SRT (Ahn et al., 2024a), often depend on proprietary models to generate training data, require large-scale preference datasets (e.g., 17k pairs), and focus mainly on text-level alignment rather than visual grounding. Other methods, like Temporal Preference Optimization (TPO) (Li et al., 2025b), target only one dimension, such as temporal reasoning, using up to 10k pairs and relying on intermediate captioning, while overlooking spatial and cross-frame aspects. This highlights the need for a unified framework that efficiently addresses all three key failure modes without relying on costly dependencies. VideoPASTA addresses this gap by generating just 7k carefully designed preference pairs that explicitly challenge a model's spatial, temporal, and cross-frame understanding. This "quality over quantity" approach avoids the need for human annotations, captions, or proprietary models, delivering a stronger and more efficient learning signal for comprehensive Video-LLM alignment.

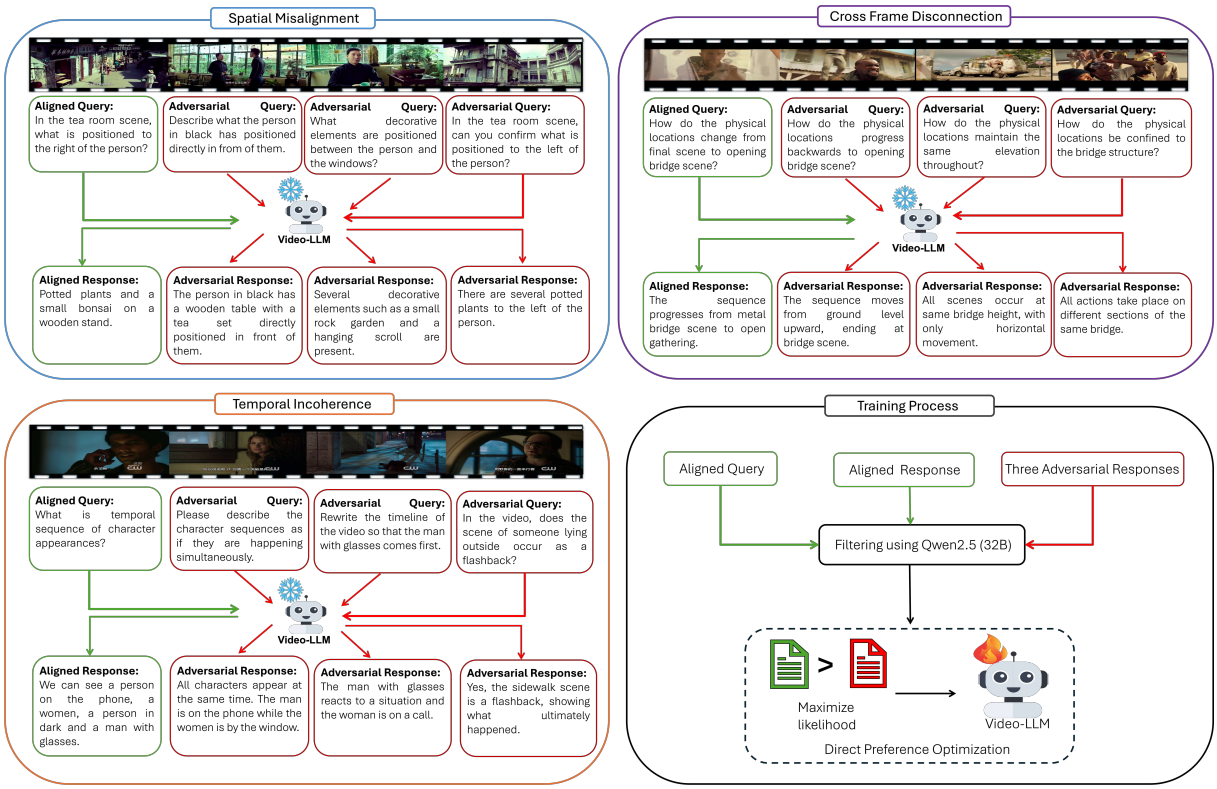


Figure 2: **Overview of VideoPASTA**. For each aligned query, we generate three types of targeted adversarial examples: (1) **Spatial Misalignment**, which intentionally distorts object positions or relationships (e.g., misplacing the plants relative to the person); (2) **Temporal Incoherence**, which violates event order (e.g., describing sequential actions as occurring simultaneously); and (3) **Cross-Frame Disconnection**, which introduces incorrect links across distant frames (e.g., misrepresenting location changes). We filter these pairs using Qwen2.5-32B (Yang et al., 2024) to ensure quality and use them to train the model via DPO, optimizing for a larger likelihood gap between aligned and adversarial responses. 🔥 → trainable, ❄ → frozen.

3 VideoPASTA

VideoPASTA is a DPO-based framework designed to improve Video-LLM alignment by addressing three key failure modes: spatial misalignment, temporal incoherence, and cross-frame disconnection. It optimizes over a structured preference dataset $D = \{(V, q, r^+, r^-)\}$, where V is the input video, q is a query, r^+ is the aligned (preferred) response, and r^- is an adversarial (misleading) response that introduces deliberate misalignment.

For each input video, we generate an aligned query q. We then produce a aligned response r^+ using densely sampled frames (32fps) and an adversarial response r^- using sparsely sampled frames (1fps), elicited by a deliberately flawed *adversarial query*. This adversarial query is only a data generation tool; the final training triplet remains (V, q, r^+, r^-), pairing the adversarial response with the original aligned query. The 32:1 sampling ratio is a design choice to ensure factual accuracy in aligned responses while inducing errors in adversarial ones. Training on these pairs via DPO improves the model's video-language alignment. Figure 2 shows the full pipeline. All prompts are provided in the Appendix (Figures 9, 10, 11, and 12).

3.1 Spatial Misalignment

We create targeted preference pairs that focus on spatial alignment.

Query Generation. We generate a variety of spatial queries covering key aspects of spatial understanding, including occlusion (e.g., "Which object is partially hidden?"), depth perception (e.g., "Which item appears closest to the camera?"), relative positioning (e.g., "How many objects are present in the left third vs. the right third of the frame?"), foreground-background relationships, and frame layout (e.g., "Which objects are near the top edge vs. the bottom edge?").

Response Generation. We generate aligned responses from videos sampled at their native frame rate to capture fine-grained spatial details. This

ensures that the generated aligned responses accurately reflect the true spatial relationships in the video. Corresponding adversarial responses are generated using prompts specifically designed to induce spatial errors (e.g., describing occluded objects as fully visible or ignoring depth cues).

3.2 Temporal Incoherence

To enhance temporal reasoning, we create preference pairs that focus on the model's temporal coherence.

Query Generation. We generate queries focused on key temporal aspects, including event ordering (e.g., "What occurs first?"), action boundaries (e.g., "Does the person complete one task before starting the next?"), transition points (e.g., "When does the subject switch activities?"), and causality (e.g., "Is the second event a direct result of the first?").

Response Generation. We generate aligned responses that accurately describe the sequence of events, effectively capturing transitions, dependencies, and causal relationships. Corresponding adversarial responses are generated using prompts that induce temporal distortions (e.g., describing sequential actions as simultaneous or merging distinct events).

3.3 Cross-Frame Disconnection

Robust video understanding requires capturing long-range cross-frame relationships. We create targeted preference pairs that focus on these dependencies.

Query Generation. We generate queries focused on cross-frame dependencies, including object continuity (e.g., "Does the same object reappear in both the opening and closing scenes?") and narrative links (e.g., "Do early events foreshadow later developments?").

Response Generation. We generate aligned responses that accurately reflect object transformations, character continuity, setting changes, and narrative flow. Adversarial responses are generated using prompts that intentionally break these continuities (e.g., describing "a new red car appears" when it is the same vehicle from a different angle).

3.4 Preference Data Filtering

We generate three adversarial examples for each failure mode per query and use an open-source LLM, Qwen2.5-32B (Yang et al., 2024) as a lightweight verification step to ensure the adversarial examples are genuinely incorrect. We prompt Qwen2.5-32B with a textual comparison task to verify that each adversarial example introduces a deliberate misalignment rather than simply rephrasing the correct answer. Adversaries that are too similar to the aligned samples or lack clear contradictions are discarded and regenerated. Similarly, we perform a "sanity check" on aligned responses to ensure they correctly align with the queries without errors. This filtering process creates preference pairs that accurately represent the targeted failure modes, enabling more precise alignment during DPO.

3.5 Training Process

VideoPASTA leverages structured preference pairs to address distinct failure modes in video understanding through DPO. We begin by partitioning the preference dataset $\mathcal{D} = \{(V, q, r^+, r^-)\}$ into three subsets: $\mathcal{D}_s$, $\mathcal{D}_t$, and $\mathcal{D}_c$, corresponding to spatial, temporal, and cross-frame alignment, respectively. For a video-language model M_θ, we define the DPO loss for a single preference pair as:

$$\begin{aligned} \Delta(V, q, r^+, r^-) &= \log p_\theta(r^+ \mid V, q) \\ &\quad - \log p_\theta(r^- \mid V, q), \\ \mathcal{L}_{\text{DPO}}(V, q, r^+, r^-) &= -\log \sigma\Big(\lambda\, \Delta(V, q, r^+, r^-)\Big), \end{aligned} \tag{1}$$

where σ is the sigmoid function and λ is a scaling factor. We then compute the DPO loss for each subset of preference pairs, weighted by α, β, and γ for spatial, temporal, and cross-frame alignment, respectively. The overall training objective is:

$$\begin{aligned} \mathcal{L} = \alpha\, \mathbb{E}_{\mathcal{D}_s}\big[\mathcal{L}_{\text{DPO}}\big] + \beta\, \mathbb{E}_{\mathcal{D}_t}\big[\mathcal{L}_{\text{DPO}}\big] \\ + \gamma\, \mathbb{E}_{\mathcal{D}_c}\big[\mathcal{L}_{\text{DPO}}\big]. \end{aligned} \tag{2}$$

This formulation allows us to adjust the model's focus on different aspects of video-language alignment during training.

4 Experiments and Evaluation

For training, we sample 3,000 videos from ActivityNet (Yu et al., 2019), whose diversity of *203 complex activities* provides a strong foundation for learning the *fundamental reasoning skills* our framework targets. This dataset is not part of our evaluation benchmarks, ensuring our results reflect true generalization. We use a large model (InternVL2.5-38B) to generate queries but all aligned and adversarial responses are generated by the smaller target models themselves. This

Model	TempCompass (Avg.)	PerceptionTest (val_mc)	NeXTQA (mc_test)	MVBench	MLVU (dev)	LongVideoBench (val_v)	VideoMME (w/o sub)
State-of-the-Art Models							
VideoLLaMA2† (Cheng et al., 2024)	43.4	51.4	-	54.6	35.5	-	47.9
Kangaroo† (Liu et al., 2024a)	-	-	-	61.0	61.0	54.8	56.0
LLaVA-NeXT-Video† (Zhang et al., 2024c)	53.0	48.8	53.5	53.1	-	49.1	46.5
LongVA† (Zhang et al., 2024b)	-	-	-	-	58.8	51.3	52.6
Qwen2-VL (Wang et al., 2024b)	68.9	62.3	75.7	64.9	57.5	55.6	55.3
LLaVA-Video (Zhang et al., 2024d)	66.4	67.9	74.2	58.6	66.5	58.2	62.4
Off-the-Shelf Preference-Optimized Models							
LLaVA-Hound-DPO (Zhang et al., 2025)	55.5	45.1	61.6	36.6	41.1	36.7	34.2
i-SRT (Ahn et al., 2024a)	56.0	47.0	63.0	36.3	39.9	38.2	34.7
LLaVA-Video-TPO (Li et al., 2025b)	66.6	66.3	77.8	56.7	66.3	58.3	62.4
Model-Agnostic Preference Optimization using VideoPASTA							
LLaVA-NeXT-Interleave (Baseline)	54.1	51.2	67.0	46.5	52.5	44.8	48.3
+ SFT	54.3	51.5	67.4	46.8	52.7	45.0	48.5
+ Hound-DPO (Zhang et al., 2025)	51.7 (−2.4)	49.5 (−1.7)	65.8 (−1.2)	44.3 (−2.2)	50.2 (−2.3)	42.5 (−2.3)	46.7 (−1.6)
+ TPO (Li et al., 2025b)	54.3 (+0.2)	52.4 (+1.2)	68.5 (+1.5)	47.9 (+1.4)	53.6 (+1.1)	46.5 (+1.7)	49.6 (+1.3)
+ **VideoPASTA**	**56.4** (+2.3)	**53.8** (+2.6)	**70.1** (+3.1)	**49.0** (+2.5)	**55.8** (+3.3)	**47.9** (+3.1)	**51.4** (+3.1)
LLaVA-OneVision (Baseline)	64.5	57.1	79.3	56.7	64.9	56.3	58.2
+ SFT	64.6	57.4	79.3	56.9	65.1	56.5	58.1
+ Hound-DPO (Zhang et al., 2025)	63.2 (−1.3)	55.8 (−1.3)	78.1 (−1.2)	55.3 (−1.4)	63.2 (−1.7)	54.8 (−1.5)	56.9 (−1.3)
+ TPO (Li et al., 2025b)	65.6 (+1.1)	58.4 (+1.3)	80.6 (+1.3)	57.9 (+1.2)	65.8 (+0.9)	57.5 (+1.2)	59.2 (+1.0)
+ **VideoPASTA**	**67.2** (+2.7)	**60.3** (+3.2)	**81.8** (+2.5)	**59.1** (+2.4)	**67.5** (+2.6)	**58.5** (+2.2)	**60.1** (+1.9)
InternVL2.5 (Baseline)	68.3	62.2	77.0	69.8	59.5	52.9	57.9
+ SFT	68.2	62.3	77.4	70.4	59.4	53.0	58.1
+ Hound-DPO (Zhang et al., 2025)	66.8 (−1.5)	61.0 (−1.2)	74.8 (−2.2)	64.2 (−5.6)	60.2 (+0.7)	54.3 (+1.4)	54.6 (−3.3)
+ TPO (Li et al., 2025b)	68.2 (−0.1)	62.0 (−0.2)	77.2 (+0.2)	68.8 (−1.0)	61.5 (+2.0)	58.1 (+5.2)	60.0 (+2.1)
+ **VideoPASTA**	**71.9** (+3.6)	**66.1** (+3.9)	**80.7** (+3.7)	**73.8** (+4.0)	**63.4** (+3.9)	**58.1** (+5.2)	**62.0** (+4.1)
Qwen2.5-VL (Baseline)	71.7	68.6	75.8	65.2	68.7	60.7	62.2
+ SFT	71.8	69.1	77.2	65.5	68.8	60.9	62.5
+ Hound-DPO (Zhang et al., 2025)	70.3 (−1.4)	67.6 (−1.0)	76.1 (+0.3)	65.7 (+0.5)	66.4 (−2.3)	56.3 (−4.4)	63.2 (+1.0)
+ TPO (Li et al., 2025b)	71.5 (−0.2)	69.0 (+0.4)	**77.6** (+1.8)	65.3 (+0.1)	68.9 (+0.2)	59.2 (−1.5)	**64.2** (+2.0)
+ **VideoPASTA**	**72.3** (+0.6)	**69.4** (+0.8)	77.3 (+1.5)	**66.3** (+1.1)	**69.2** (+0.5)	**61.5** (+0.8)	64.1 (+1.9)

Table 1: **Comprehensive evaluation of VideoPASTA against leading (7B) video understanding models.** The best scores are in **bold**, and the second-best scores are underlined. Results marked with † are from the original papers; all other results are reproduced using LMMs-Eval (Zhang et al., 2024a).

setup ensures models learn to refine their own outputs rather than simply distilling knowledge from a more capable model. Our structured adversarial sampling pipeline initially generates 90,000 preference pairs, which after rigorous filtering (details in Appendix §E.1), are reduced to 7,020 high-quality pairs. We fine-tune models using the SWIFT (Zhao et al., 2025) framework for efficient adaptation. All training and evaluations are performed on four NVIDIA L40S GPUs (48GB each), with a maximum input of 32 frames per video to prevent CUDA out-of-memory errors. We employ LoRA (Hu et al., 2022) with rank $r = 8$ and $\alpha_{LoRA} = 8$. For DPO, we set the scaling factor λ to 0.1. The overall training loss (Eq.2) combines three components: spatial ($\alpha_S = 0.4$), temporal ($\beta_T = 0.4$), and cross-frame ($\gamma_C = 0.2$) alignment. We apply VideoPASTA to four diverse foundation models: Qwen2.5-VL (7B) (Bai et al., 2025), LLaVA-NeXT-Interleave (7B) (Li et al., 2025a), LLaVA-OneVision (7B) (Li et al., 2024a), and InternVL2.5 (8B) (Chen et al., 2024c). To ensure fair model-agnostic comparisons, we limit training data to 7,020 high-quality preference pairs for all models. This corresponds to the smallest filtered set (from LLaVA-NeXT-Interleave), with larger sets from other models subsampled accordingly to maintain consistency in data quantity. Evaluation is conducted using LMMs-Eval (Zhang et al., 2024a) to ensure fair comparisons with prior work.

The Appendix contains additional experiments and information, including: DPO training dynamics (§A), preference learning on smaller (1B-3B) models (§B), adversarial robustness of VideoPASTA (§C), Qwen2.5-VL-specific ablations (§D), full dataset statistics and adversarial samples (§E), qualitative examples (§F), and all prompt templates (§G).

Benchmarks. We evaluate on general video understanding benchmarks: TempCompass (Liu et al., 2024b) (temporal understanding), PerceptionTest (Patraucean et al., 2024) (visual percep-

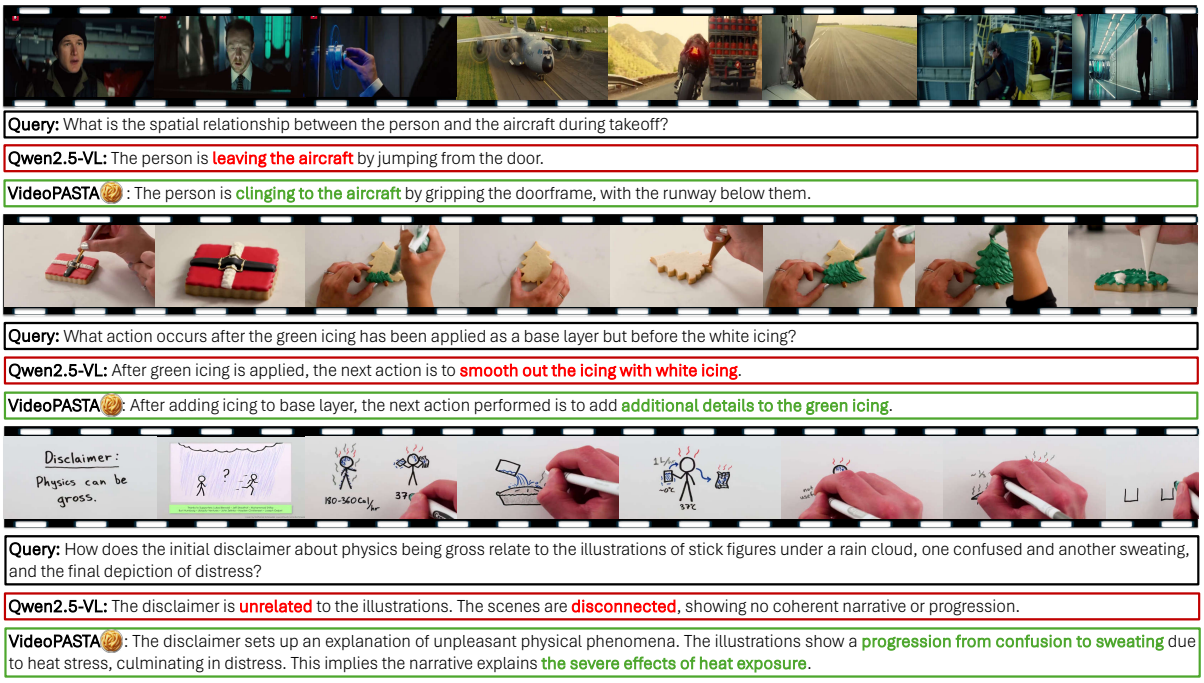


Figure 3: **Qualitative comparison of VideoPASTA against Qwen2.5-VL** (Bai et al., 2025). Examples show VideoPASTA improves (1) *Spatial reasoning* (aircraft interaction), (2) *Temporal understanding* (icing sequence), and (3) *Cross-frame reasoning* (narrative connection in stick figures), where the baseline fails.

tion), NeXTQA (Xiao et al., 2021) (compositional reasoning), and MVBench (Li et al., 2024b) (multi-task reasoning). For long-form evaluation, we use LongVideoBench (Wu et al., 2024) (hour-long videos), MLVU (Zhou et al., 2025) (multi-task, 3-minute to 2-hour videos), and VideoMME (Fu et al., 2025) (6 visual domains, 30 subfields, 11-second to 1-hour videos).

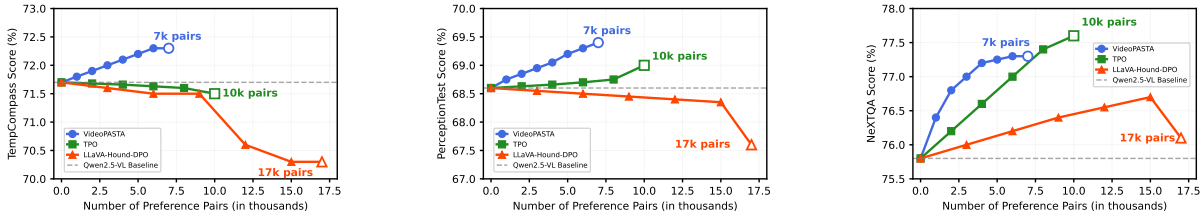
4.1 Results

We compare VideoPASTA with (1) the original foundation models listed above, (2) other state-of-the-art models, and (3) off-the-shelf models enhanced via preference optimization. Table 1 presents the evaluation results. Figure 3 shows qualitative examples of how VideoPASTA improves spatial, temporal, and cross-frame reasoning.

VideoPASTA enhances all foundation models. VideoPASTA performs well across various foundation models, showing strong generalizability and consistent performance improvements. For instance, VideoPASTA combined with Qwen2.5-VL achieves top scores on TempCompass, PerceptionTest, MLVU, and LongVideoBench. Similarly, VideoPASTA with LLaVA-OneVision attains the highest score on NeXTQA. In addition, VideoPASTA with InternVL2.5 achieves the best result on MVBench and improves the VideoMME score by +4.1 percentage points. These results show that VideoPASTA's targeted alignment helps a wide range of Video-LLMs. In contrast, simple supervised fine-tuning (SFT) using only aligned responses leads to only small improvements. This highlights the importance of training with adversarial preferences through DPO.

Comparison with State-of-the-Art. Compared to other state-of-the-art models listed in Table 1, VideoPASTA combined with Qwen2.5VL outperforms all models on all benchmarks, surpassing strong baselines like Qwen2-VL and LLaVA-Video. Key improvements include a +5.9 percentage point gain in temporal reasoning on TempCompass and a +1.5 increase on PerceptionTest over LLaVA-Video, which is instruction-tuned on 1.3M SFT pairs. VideoPASTA also shows strong performance on long-form video tasks, achieving +3.3 percentage point gain on LongVideoBench, +2.7 on MLVU, and +1.7 on VideoMME compared to LLaVA-Video. Similarly, when paired with the other three foundation models, VideoPASTA outperforms SOTA models on several, though not all, benchmarks. These results highlight VideoPASTA's ability to elevate smaller models to SOTA performance through targeted and data-efficient training.

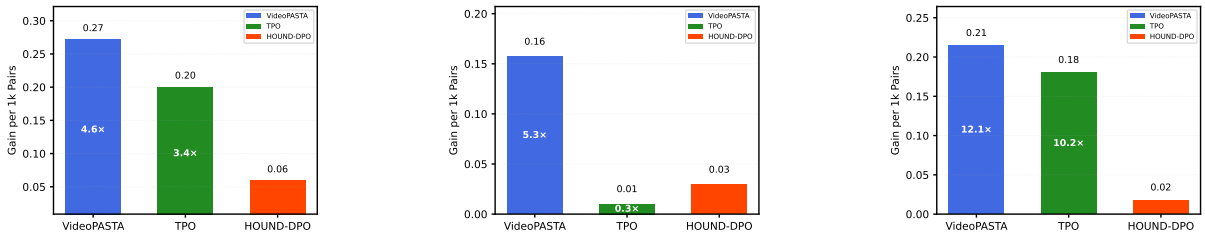
Outperforming Preference-Optimized Models. Compared to preference-optimized models, VideoPASTA combined with Qwen2.5VL outper-



(a) TempCompass (b) PerceptionTest (c) NeXTQA

(d) VideoMME (e) MLVU (f) LongVideoBench

Figure 4: Performance vs. # of Preference Pairs across six benchmarks. **VideoPASTA achieves superior results with only 7k pairs** compared to TPO (Li et al., 2025b) (10k pairs) and Hound-DPO (Zhang et al., 2025) (17k pairs).



(a) VideoMME (b) MVBench (c) NeXTQA

Figure 5: **Information gain analysis across three representative benchmarks**. Each bar represents performance improvement per 1k preference pairs, calculated as (final score - baseline score) / # of pairs in thousands.

forms LLaVA-Hound-DPO (Zhang et al., 2025) and i-SRT (Ahn et al., 2024a) on all seven benchmarks and surpasses LLaVA-Video-TPO (Li et al., 2025b) on six out of seven benchmarks (except NeXTQA) by a significant margin. On the other hand, VideoPASTA paired with LLaVA-OneVision and InternVL2.5 outperforms LLaVA-Video-TPO on NeXTQA. These improvements highlight that our multi-dimensional approach, which tackles critical failure modes in Video-LLMs, addresses fundamental challenges in prior work while requiring minimal resources.

4.2 Ablation Studies

How efficient is VideoPASTA compared to other preference optimization approaches? As shown in Figure 4, VideoPASTA outperforms TPO (Li et al., 2025b) (10k pairs, 96 frame sampling) and LLaVA-Hound-DPO (Zhang et al., 2025) (17k pairs) using only 7k preference pairs, highlighting its significantly higher efficiency. In addition, VideoPASTA maintains stable performance as the number of training pairs increases, whereas other methods occasionally show performance drops on certain benchmarks with more data. This focus on "quality over quantity" is further highlighted by the information gain per 1k pairs, shown in Figure 5. For instance, on MVBench, VideoPASTA is about 16× more efficient than TPO and 5.3× more efficient than Hound-DPO. These efficiency gains translate into tangible performance improvements: when paired with InternVL2.5, VideoPASTA boosts MVBench score by +4.0 percentage points, while Hound-DPO leads to a substantial -5.6 drop, and TPO results in a -1.0 decrease. These results show that our structured adversarial examples targeting specific failure modes create a more robust learning signal than larger, more generic, or proprietary model-dependent preference datasets.

Can VideoPASTA enable self-improvement using queries and responses generated by the target model itself? Table 2(a) shows that VideoPASTA enables self-improvement even with-

Target Model	TempCompass	Perception Test	NeXTQA	MVBench	MLVU	LongVideoBench	VideoMME
(a) *VideoPASTA with Self-Generated Preferences*							
Qwen2.5-VL-7B	71.2 (−0.5)	68.9 (+0.3)	76.2 (+0.4)	65.6 (+0.4)	68.8 (+0.1)	60.9 (+0.2)	63.0 (+0.8)
InternVL2.5-8B	69.2 (+0.9)	63.4 (+1.2)	77.5 (+0.5)	70.1 (+0.3)	60.0 (+0.5)	53.7 (+0.8)	58.5 (+0.6)
(b) VideoPASTA with Preferences from Auxiliary Models							
Qwen2.5-VL-7B *Query:* InternVL2.5-38B *Response:* InternVL2.5-8B	71.9 (+0.2)	69.1 (+0.5)	76.7 (+0.9)	65.9 (+0.7)	69.0 (+0.3)	61.2 (+0.5)	63.5 (+1.3)
InternVL2.5-8B *Query:* InternVL2.5-38B *Response:* Qwen2.5-VL-7B	70.2 (+1.9)	64.5 (+2.3)	78.4 (+1.4)	71.5 (+1.7)	62.3 (+2.8)	55.6 (+2.7)	60.8 (+2.9)

Table 2: **Impact of the source of preference generation.** (a) The target model generates preference pairs without relying on any external auxiliary models. (b) Two auxiliary models are used to generate query-response (preference) pairs for DPO training the target model. Performance changes (+ / −) are relative to respective DPO target model baselines in Table 1.

Method	# of Pref.	Temporal	Spatial	Action	Object
Baseline Performance					
Qwen2.5-VL	-	35.0	63.6	54.0	55.9
Baseline Preference Datasets					
w/ TPO	10k	**47.5** (+12.5)	65.1 (+1.5)	54.3 (+0.3)	54.8 (−1.1)
w/ Hound-DPO	17k	42.2 (+7.2)	61.7 (−1.9)	52.1 (−1.9)	55.5 (−0.4)
VideoPASTA: Targeting Single Failure Modes					
VideoPASTA (Temporal Only)	2.3k	44.0 (+9.0)	67.1 (+3.5)	49.2 (−4.8)	51.3 (−4.6)
VideoPASTA (Spatial Only)	2.3k	40.1 (+5.1)	74.8 (+11.2)	55.0 (+1.0)	55.4 (−0.5)
VideoPASTA (Cross-Frame Only)	2.3k	43.3 (+8.3)	66.8 (+3.2)	54.9 (+0.9)	57.2 (+1.3)
VideoPASTA	7k	45.2 (+10.2)	**78.6** (+15.0)	**56.1** (+2.1)	**58.4** (+2.5)

Table 3: **Effect of targeted failure modes on VideoMME tasks.** Gains/losses relative to baseline Qwen2.5-VL.

out an external auxiliary model. When target models like Qwen2.5-VL-7B or InternVL2.5-8B generate their *own* query-response (preference) pairs for DPO alignment, they still achieve consistent performance improvements. For example, Qwen2.5-VL using self-generated queries and responses improves by +0.4 percentage points on MVBench and +0.8 on VideoMME compared to its baseline. This result is significant as it demonstrates a fully self-improving loop where no external 'teacher' model is used. The consistent gains (e.g., +0.8 on VideoMME for Qwen2.5-VL) confirm that the framework's effectiveness stems from our targeted alignment methodology itself, proving that the model is learning to correct its own failures rather than merely distilling knowledge.

Can VideoPASTA's preference signals generalize when the query-response generators and the DPO target model are different? Table 2(b) shows that VideoPASTA's preference signals generalize well, even when the query-response generators differ from the target model. For instance, InternVL2.5-8B improves by +2.9 percentage points on VideoMME and +2.8 on MLVU when trained on preferences generated by InternVL-38B and Qwen2.5-VL-7B. This suggests that the target model does not rely on the generation style of any specific model, proving it learns transferable rules about video understanding rather than simply memorizing the patterns of one particular preference generator.

What is the advantage of VideoPASTA's three-dimensional adversarial preferences compared to narrower or more generic preference data generation approaches? Table 3 shows that each failure mode (spatial, temporal, and cross-frame) in our adversarial sampling pipeline provides distinct benefits. Training only with temporal samples greatly improves temporal reasoning (+9.0 percentage points) but harms action (-4.8) and object (-4.6) reasoning. Similarly, training exclusively on spatial samples boosts spatial reasoning the most (+11.2) but reduces object reasoning. Combining all three failure modes delivers the best overall performance, with substantial improvements in temporal (+10.2), spatial (+15.0), action (+2.1), and object (+2.5) reasoning. In contrast, applying existing preference data from TPO or LLaVA-Hound-DPO shows suboptimal effects. TPO improves temporal reasoning (+12.5) but worsens object reasoning, while Hound-DPO's data significantly reduces spatial and action scores. These results confirm that adversarial sampling targeting multiple failure modes provides complementary benefits, leading to more comprehensive video-language alignment than relying on any single mode or generic preference data alone.

How well does VideoPASTA generalize to unseen and challenging video domains? To validate the generalization of our adversarial alignment,

Model	MovieChat	EgoSchema
LLaVA-NeXT-Interleave		
Baseline	40.0	51.0
+ **VideoPASTA**	**41.1** (+1.1)	**51.9** (+0.9)
LLaVA-OneVision		
Baseline	44.0	64.0
+ **VideoPASTA**	**45.6** (+1.6)	**65.0** (+1.0)
InternVL2.5		
Baseline	46.8	52.0
+ **VideoPASTA**	**47.4** (+0.6)	**53.6** (+1.6)
Qwen2.5-VL		
Baseline	44.2	57.6
+ **VideoPASTA**	**46.1** (+1.9)	**58.1** (+0.5)

Table 4: **Cross-Domain Generalization to Unseen Domains.** VideoPASTA demonstrates consistent performance gains on movie and egocentric video benchmarks.

we evaluate VideoPASTA on challenging, unseen video domains. As shown in Table 4, VideoPASTA consistently improves performance across all foundation models, boosting scores by up to +1.9 percentage points on MovieChat (Song et al., 2024) and +1.6 on EgoSchema (Mangalam et al., 2023). Notably, these gains are achieved even though training occurs exclusively on ActivityNet. This result shows that by addressing fundamental reasoning failures in spatial, temporal, and cross-frame dimensions, we foster more robust generalization, allowing the model to adapt to diverse video types without domain-specific fine-tuning.

How does improving high-level reasoning affect lower-level perception? A core design choice of VideoPASTA is to target three high-level reasoning failures, hypothesizing that this will also enhance lower-level perceptual abilities. We validate this by analyzing performance on the sub-tasks of the MVBench (Li et al., 2024b) benchmark. As shown in Table 5, our approach improves performance on 8 out of 9 action and attribute sub-tasks. VideoPASTA achieves substantial gains of +12.0 percentage points in Action Localization and +7.0 in Action Count. These results provide strong empirical evidence that our strategy of correcting fundamental reasoning failures leads to a more holistic alignment, enhancing not just abstract understanding but also the model's core perceptual capabilities.

How well do adversarial examples target their intended failure modes? We validate the accuracy of our adversarial data generation using GPT-4o (prompt in Appendix, Figure 15) to judge whether 200 randomly sampled adversarial examples correctly induced their intended failure mode. As detailed in Table 6, our method achieves high targeting accuracy: 96.1% for spatial misalignment, 92.4% for temporal incoherence, and 88.3% for cross-frame disconnection. This result confirms that our prompt-based strategy effectively generates varied and targeted examples that provide a strong learning signal for DPO.

MVBench Task	Qwen2.5-VL	Qwen2.5-VL + VideoPASTA	Improvement
Action-Related Tasks			
Action Sequence	79.5	**80.5**	+1.0
Action Prediction	67.5	67.0	-0.5
Action Antonym	87.0	**89.5**	+2.5
Fine-grained Action	49.5	**52.5**	+3.0
Unexpected Action	80.5	**82.5**	+2.0
Action Localization	55.0	**67.0**	+12.0
Action Count	46.5	**53.5**	+7.0
Attribute-Related Tasks			
State Change	58.0	**62.0**	+4.0
Moving Attribute	90.5	**92.0**	+1.5

Table 5: **MVBench Sub-task Performance Breakdown.** Improvements on lower-level perceptual tasks support our design rationale of targeting high-level reasoning failures.

Failure Mode	Targeting Accuracy (%)
Spatial Misalignment	96.1
Temporal Incoherence	92.4
Cross-Frame Disconnection	88.3
Average	**92.3**

Table 6: **Failure Mode Targeting Accuracy by Category.**

5 Conclusion

We introduce VideoPASTA, a DPO-based framework that improves Video-LLMs via structured adversarial sampling targeting spatial, temporal, and cross-frame misalignments. Using just 7,020 preference pairs, without human supervision or video captions, our model-agnostic approach achieves significant gains across seven benchmarks with efficient 32-frame sampling. VideoPASTA demonstrates that targeted adversarial examples enable more effective learning than generic instruction tuning. While individual failure modes enhance specific capabilities, combining all three leads to broader and more comprehensive video understanding. This strategy reduces reliance on large-scale datasets, promoting resource-efficient video-language alignment. Future work can explore better evaluation metrics for model reasoning.

Limitations

While VideoPASTA demonstrates significant advancements, certain aspects warrant future exploration. The quality and diversity of the generated preference pairs depend on the capabilities of the models used in our pipeline (i.e., query generator, response generator, and verifier). Potential biases or limitations in these foundational models could subtly affect the preference dataset.

Acknowledgments

This research was supported by the National Eye Institute (NEI) of the National Institutes of Health (NIH) under award number R01EY034562. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Daechul Ahn, Yura Choi, San Kim, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024a. i-srt: Aligning large multimodal models for videos by iterative self-retrospective judgment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024b. Tuning Large Multimodal Models for Videos using Reinforcement Learning from AI Feedback. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 923–940.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, and et al. 2024a. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18407–18418.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024b. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. 2024. Vidhal: Benchmarking temporal hallucinations in vision llms. *arXiv preprint arXiv:2411.16771*.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems (NeurIPS)*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 24108–24118.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 18135–18143.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14271–14280.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Yogesh Kulkarni and Pooyan Fazli. 2025a. Avatar: Reinforcement learning to see, hear, and reason over video. *arXiv preprint arXiv:2508.03100*.

Yogesh Kulkarni and Pooyan Fazli. 2025b. Videosavi: Self-aligned video language models without human supervision. *Conference on Language Modeling (COLM)*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, and 1 others. 2024. Mitigating object hallucinations

in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2025a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *International Conference on Learning Representations (ICLR)*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.

Rui Li, Xiaohan Wang, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. 2025b. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, and 1 others. 2024c. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

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 (EMNLP)*, pages 5971–5984.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024a. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

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)*, pages 8731–8772.

Xinyu Ma, Yifan Li, Hao Wang, and 1 others. 2024. Vista-llama: Reducing hallucination in video language models via equal distance attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. EgoSchema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2024. Gpt-4o. https://openai.com/index/hello-gpt-4o/.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2024. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. 2024. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12987.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and 1 others. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *International Conference on Machine Learning (ICML)*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, and et al. 2024. Aligning Large Multimodal Models with Factually Augmented RLHF. In *Findings of the Association for Computational Linguistics (ACL)*, pages 13088–13110.

Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, and et al. 2024. Koala: Key frame-conditioned long video-llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13581–13591.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. Mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8078–8088.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. 2024c. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and 1 others. 2024d. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision (ECCV)*, pages 396–416.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, and 1 others. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2025. Vca: Video curious agent for long video understanding. *International Conference on Computer Vision (ICCV)*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*.

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*.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, and Yiming Yang. 2025. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024c. Llava-next: A strong zero-shot video understanding model.

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

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, and 1 others. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 13691–13701.

Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. 2025. Video-star: Self-training enables video instruction tuning with any supervision. *International Conference on Learning Representations (ICLR)*.

Appendix

A DPO Training Dynamics

Figure 6 shows the DPO training process for Qwen2.5-VL. The model quickly learns to distinguish between aligned responses (r^+) and adversarial ones (r^-), as shown by the growing gap between the chosen rewards (green, increasing) and the rejected rewards (red, generally decreasing). At the same time, reward accuracy (blue) rises rapidly and stabilizes around 70-75%, indicating a consistent preference for well-grounded responses over those with targeted misalignments. This demonstrates the effectiveness of our DPO-based alignment approach.

B Preference Learning with VideoPASTA on Small Models

To further assess the broad applicability and efficiency of VideoPASTA, we evaluate its performance on a range of smaller foundational models, with parameters varying from 1B to 3B. The results, presented in Table 7, demonstrate that VideoPASTA consistently provides performance uplifts even for these more compact architectures. For instance, when applied to Qwen2-VL (2B), VideoPASTA improves scores across all seven benchmarks, such as a +1.7 percentage point gain on MVBench (from 60.8 to 62.5) and +1.1 on VideoMME (from 50.1 to 51.2). Similarly, InternVL (1B) + VideoPASTA sees gains like +1.3 on MVBench and +0.5 on VideoMME.

These consistent improvements on smaller models highlight several advantages of VideoPASTA's targeted adversarial alignment. Firstly, it highlights that our novel data curation strategy, focusing on specific failure modes (spatial, temporal, cross-frame), provides a learning signal that is effective even for models with lower capacity. Secondly, the ability to boost these smaller models demonstrates that VideoPASTA is not solely reliant on the extensive pre-existing knowledge of very large foundation models to achieve its gains but can instill more robust visual reasoning directly. This reinforces the idea that the 7k targeted preference pairs efficiently address core weaknesses, offering a resource-friendly path to enhancing Video-LLMs, making video understanding capabilities more accessible.

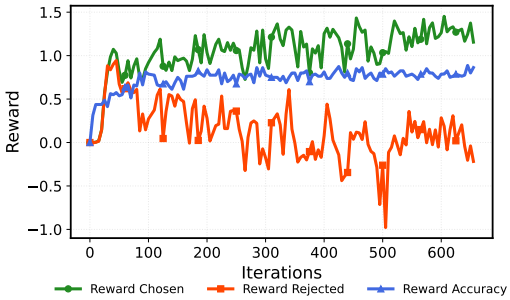


Figure 6: **DPO training converges on well-grounded responses.**

C Adversarial Robustness

To evaluate VideoPASTA's robustness against failure modes, we test 100 videos from LLaVA-Video (Zhang et al., 2024d) using GPT-4o (OpenAI, 2024) (prompt provided in Appendix, Figure 13) to generate both adversarial questions (unanswerable queries) and adversarial options (where "None of the Above" is correct) per failure mode. As shown in Table 8, VideoPASTA significantly outperforms baselines across all categories, with the most substantial gains in temporal reasoning (+14.5 percentage points). This improved robustness stems directly from our training approach, by exposing the model to targeted adversarial examples during preference optimization, VideoPASTA learns to recognize and reject similar misleading inputs during inference. Unlike generic preference optimization, our structured adversarial sampling creates a more discriminative model capable of identifying spatial inconsistencies, temporal contradictions, and cross-frame disconnections. GPT-4o was also used to evaluate model responses (prompt provided in Appendix, Figure 14), specifically identifying rejection phrases like "cannot be answered" and "insufficient information" when models correctly recognized adversarial inputs.

D Qwen2.5-VL Specific Ablations

The following ablations are performed using Qwen2.5-VL as the target model to understand the impact of specific hyperparameter choices within the VideoPASTA framework when applied to this backbone:

D.1 DPO Weight Ratio Analysis

We explore the impact of different weighting schemes for the DPO loss components (spatial:α, temporal:β, cross-frame:γ) on Qwen2.5-VL, de-

Model	TempCompass	PerceptionTest	NeXTQA	MVBench	MLVU	LongVideoBench	VideoMME
Qwen2-VL (2B)	62.0	53.0	69.1	60.8	51.3	46.6	50.1
+ **VideoPASTA**	63.6 (+1.6)	54.4 (+1.4)	70.6 (+1.5)	62.5 (+1.7)	51.9 (+0.6)	47.7 (+1.1)	51.2 (+1.1)
Qwen2.5-VL (3B)	66.6	63.1	74.9	63.5	65.9	55.8	61.2
+ **VideoPASTA**	67.3 (+0.7)	63.9 (+0.8)	75.6 (+0.7)	64.5 (+1.0)	66.7 (+0.8)	56.0 (+0.2)	61.8 (+0.6)
InternVL2.5 (1B)	41.6	55.0	65.3	63.5	55.5	45.4	49.5
+ **VideoPASTA**	42.5 (+0.9)	55.7 (+0.7)	66.4 (+1.1)	64.8 (+1.3)	56.1 (+0.6)	45.7 (+0.3)	50.0 (+0.5)
InternVL2.5 (2B)	47.0	57.3	68.4	65.9	56.2	48.0	53.0
+ **VideoPASTA**	48.2 (+1.2)	58.1 (+0.8)	69.4 (+1.0)	67.3 (+1.4)	56.9 (+0.7)	48.5 (+0.5)	54.3 (+1.3)

Table 7: **Preference Learning with VideoPASTA on small models**.

Model	Spatial Misalignment		Temporal Incoherence		Cross-Frame Disconnection	
	Adv. Question (%)	Adv. Options (%)	Adv. Question (%)	Adv. Options (%)	Adv. Question (%)	Adv. Options (%)
Qwen2.5-VL (Bai et al., 2025)	38.4	42.6	35.2	39.5	31.8	36.7
LLaVA-Hound-DPO (Zhang et al., 2025)	39.2 (+0.8)	43.1 (+0.5)	36.5 (+1.3)	39.8 (+0.3)	31.9 (+0.1)	37.2 (+0.5)
TPO (Li et al., 2025b)	41.3 (+2.9)	44.5 (+1.9)	48.2 (+13.0)	51.4 (+11.9)	32.4 (+0.6)	37.5 (+0.8)
VideoPASTA	**46.8** (+8.4)	**51.1** (+8.5)	**49.7** (+14.5)	**52.8** (+13.3)	**33.1** (+1.3)	**38.2** (+1.5)

Table 8: **Performance on Adversarial QA Samples Across Different Failure Modes**. "Adv. Question": Unanswerable queries (higher rejection rate is better). "Adv. Options": Questions where "None of the Above" is correct (higher NOTA selection is better). Each cell shows correct handling (%).

tailed in Table 9. While focusing on a single dimension (e.g., a 0.6:0.2:0.2 spatial focus) shows some targeted benefits, a more balanced distribution proves superior for overall performance. Our chosen configuration of $\alpha = 0.4, \beta = 0.4, \gamma = 0.2$ ("0.4:0.4:0.2 (Ours)") consistently yielded the best results across all seven benchmarks, indicating that while spatial and temporal aspects are crucial, a non-negligible weight for cross-frame reasoning is also important for comprehensive alignment.

D.2 Number of Adversarial Examples per Aligned Sample

Table 9 shows that performance is optimal with three adversarial examples corresponding to our three targeted failure modes. Using fewer examples leaves certain aspects of video understanding insufficiently challenged, while more examples lead to diminishing returns. This confirms that pairing each aligned response with exactly three adversarial responses, one for each failure mode, best reinforces alignment across spatial, temporal, and cross-frame reasoning.

D.3 Frame Sampling

Our analysis of sampling rates (Table 9) shows that using uniformly dense sampling for both aligned and adversarial examples lowers performance as models struggle to detect subtle alignment errors. The optimal configuration (32:1) strikes a balance: dense aligned sampling captures temporal details, while sparse adversarial sampling creates clear misalignment patterns. This result is consistent across benchmarks, highlighting the importance of a well-designed sampling strategy in model training.

D.4 Image and Video Resolution Settings

Ablations on image and video resolution for Qwen2.5-VL (Table 9) confirm that higher resolutions generally contribute to better performance, with our selected settings (MAX image resolution $128 \times 28 \times 28$ and VID_MAX video resolution $64 \times 28 \times 28$) providing a strong balance for our experiments. It is worth noting that score discrepancies with the original Qwen2.5-VL paper may arise because the Qwen team utilized substantially higher input parameters (e.g., video_max_pixels up to $768 \times 28 \times 28$, max_frames up to 768), which, while potentially beneficial, are often impractical for typical computing environments and our focus on resource-efficient alignment.

E Dataset Overview

E.1 Dataset Statistics

Starting with 3000 videos from ActivityNet (Yu et al., 2019), we systematically generate preference pairs through structured adversarial sampling. For each video V, we generate 10 queries Q targeting different aspects of video understanding. Each query $q \in Q$ is paired with three targeted adversarial responses r_{spatial}, r_{temporal}, and $r_{\text{crossframe}}$, rep-

Configuration	TempCompass	Perception Test	NeXTQA	MVBench	MLVU	LongVideoBench	VideoMME
DPO Weight Ratio (α:β:γ)							
0.33:0.33:0.33 (Equal Weights)	71.5	68.2	76.8	65.7	68.5	60.6	63.4
0.6:0.2:0.2 (Spatial Focus)	71.6	69.3	76.5	65.5	68.1	60.3	63.1
0.2:0.6:0.2 (Temporal Focus)	72.2	68.1	76.3	65.4	68.7	60.8	63.2
0.2:0.2:0.6 (Cross-Frame Focus)	71.2	67.8	76.9	65.8	68.9	61.2	63.5
0.4:0.4:0.2 (Ours)	**72.3**	**69.4**	**77.3**	**66.3**	**69.2**	**61.5**	**64.1**
Adversarial Examples per Aligned Sample							
1	71.5	68.2	76.2	65.0	67.5	58.8	62.3
2	72.0	68.9	76.9	65.8	68.4	60.2	63.2
3 (Ours)	**72.3**	**69.4**	**77.3**	**66.3**	**69.2**	**61.5**	**64.1**
4	72.1	69.2	77.0	66.0	68.9	61.2	63.8
5	71.9	69.0	76.8	65.9	68.7	61.0	63.6
Frame Sampling (Aligned:Adversarial)							
32:32	71.6	68.5	76.7	65.6	68.7	60.8	62.4
16:16	71.5	68.4	76.6	65.5	68.6	60.7	62.3
32:8	72.0	69.0	77.1	66.0	69.0	61.2	63.5
16:4	71.9	68.9	77.0	65.9	68.9	61.0	63.2
32:1 (Ours)	**72.3**	**69.4**	**77.3**	**66.3**	**69.2**	**61.5**	**64.1**
16:1	72.1	69.2	77.2	66.1	69.0	61.3	63.7
Image Resolution Ablation							
MIN=4×28×28, MAX=64×28×28	70.1	67.2	75.2	64.1	67.3	59.4	62.0
MIN=4×28×28, MAX=96×28×28	71.4	68.5	76.4	65.4	68.5	60.6	63.2
MIN=4×28×28, MAX=128×28×28	**72.3**	**69.4**	**77.3**	**66.3**	**69.2**	**61.5**	**64.1**
Video Resolution Ablation							
VID_MIN=32×28×28, VID_MAX=32×28×28	70.5	67.6	75.4	64.5	67.6	59.8	62.3
VID_MIN=48×28×28, VID_MAX=48×28×28	71.6	68.7	76.5	65.6	68.6	60.8	63.5
VID_MIN=64×28×28, VID_MAX=64×28×28	**72.3**	**69.4**	**77.3**	**66.3**	**69.2**	**61.5**	**64.1**

Table 9: **Comprehensive Ablation Studies for VideoPASTA on Qwen2.5-VL.** This table details the impact of DPO weight ratios ($\alpha : \beta : \gamma$), the number of adversarial examples per aligned sample, frame sampling strategies (aligned:adversarial frames), and Qwen specific image/video resolutions. Our chosen configurations are in **bold**.

resenting spatial, temporal, and cross-frame failure modes, respectively. Theoretically, this setup yields:

$$N_{\text{potential}} = |V| \times |Q| \times |R^-| = 3000 \times 10 \times 3 = 90,000 \quad (3)$$

potential preference pairs, where $|V|$ is the number of videos, $|Q|$ is the number of queries per video, and $|R^-|$ is the number of adversarial responses per query.

However, to ensure dataset quality, we employ rigorous filtering using Qwen2.5-32B (Yang et al., 2024) verification using the prompt template given in Figure 12. Each preference pair must satisfy three criteria:

1. The aligned response should accurately reflect the video content relative to the query.
2. The adversarial response must introduce a clear, deliberate misalignment.
3. The misalignment must be specific to its targeted failure mode.

This verification process retains approximately 7.8% of the potential pairs (on LLaVA-NeXT-Interleave):

$$N_{\text{final}} = N_{\text{potential}} \times r_{\text{retention}} \approx 90000 \times 0.078 \approx 7,020. \quad (4)$$

where $r_{\text{retention}}$ is the retention rate after quality filtering. This filtered dataset provides a balanced representation across failure modes while maintaining high standards for preference pair quality. The strict filtering ensures that each adversarial example presents a genuine challenge for video-language alignment rather than simple errors or rephrasing.

E.2 Adversarial Sample Diversity

Figure 7 illustrates the diversity of adversarial examples demonstrating how VideoPASTA targets

Figure 7: **Adversarial example diversity**. Each row shows a video sequence with its corresponding query and aligned and adversarial responses. The adversarial samples cover *spatial misalignment* (counting objects in scenes), *temporal incoherence* (order of actions in cooking/preparation), and *cross-frame disconnection* (scene transitions and contextual shifts). Adversarial responses deliberately introduce specific misalignments by either negating obvious visual elements, confusing sequential ordering, or collapsing distinct temporal phases into simultaneous events, while aligned responses maintain accurate spatial-temporal alignment with the video content.

specific failure modes. These examples were carefully curated to challenge different aspects of video comprehension while maintaining clear distinctions between aligned and adversarial responses.

Spatial Misalignment. The boat counting example demonstrates our approach to spatial reasoning. While the adversarial response completely negates the presence of obvious visual elements ("no boats"), the challenge lies not in the simple presence/absence but in the precise spatial relationships ("positioned near the shore, with one slightly further out"). This forces the model to develop fine-

Figure 8: **Qualitative comparison of VideoPASTA against Qwen2.5-VL across key failure modes**. The examples demonstrate how our method addresses three critical challenges in video understanding: (1) **Spatial misalignment** (correctly describing the gradual progression of a solar eclipse and identifying spatial evidence in the Harry Potter scene), (2) **Temporal incoherence** (accurately capturing sequential emotional progressions in the athlete's reactions and proper cooking preparation steps), and (3) **Cross-frame disconnection** (maintaining narrative coherence from camera handling to photography outcomes and character emotions). Qwen2.5-VL responses exhibit typical failure patterns: misrepresenting spatial relationships, incorrectly sequencing temporal events, and failing to establish meaningful connections across frames. VideoPASTA responses demonstrate robust video-language alignment across all three dimensions.

grained spatial awareness rather than just object detection capabilities.

Temporal Incoherence. Two examples highlight our approach to temporal understanding. The cooking sequence tests precise transitional timing between steps, where the adversarial response artificially collapses distinct preparation phases into simultaneous actions. Similarly, the equipment

preparation example challenges the model's ability to distinguish between sequential and concurrent actions. These adversarial samples are particularly effective because they present plausible but incorrect temporal relationships.

Cross-Frame Disconnection. The scene transition example illustrates how we assess long-range comprehension. The adversarial response mistakenly interprets superficial visual changes, such as a close-up of a face, as significant narrative shifts, whereas the aligned response accurately identifies meaningful context transitions, like an external threat leading to an internal response. This evaluates the model's ability to track narrative progression across distant frames.

Each example undergoes thorough validation using Qwen2.5-32B (Yang et al., 2024) to ensure that adversarial responses reflect genuine misunderstandings rather than simple errors or rephrasings. This systematic approach to adversarial example generation reinforces robust video-language alignment across multiple dimensions of video understanding.

F Qualitative Examples

We present several representative examples that demonstrate how VideoPASTA improves video understanding across various scenarios. Figure 8 illustrates three key aspects of our model's capabilities in handling complex video content.

First, in the camera advertisement sequence, while Qwen2.5-VL (Bai et al., 2025) fails to recognize the narrative structure and describes it as "unrelated clips" VideoPASTA successfully captures the purposeful progression from technical camera operation to creative photography. This demonstrates how our cross-frame adversarial sampling helps the model develop a more coherent understanding of extended narratives. Next, the animated sequence with Bugs Bunny showcases VideoPASTA's enhanced ability to track emotional progression. Instead of merely detecting immediate reactions, our model recognizes the escalation from initial irritation to visible anger and, ultimately, to explosive rage. This improvement stems from our temporal incoherence adversarial sampling, which teaches the model to distinguish between simultaneous and sequential emotional states. The cooking demonstration particularly highlights the benefits of our local spatial alignment strategy. While the baseline model confuses the order of preparation steps, VideoPASTA correctly identifies the precise sequence of cleaning, coating, and frying the chilies. This accuracy in tracking procedural steps is crucial for practical applications like instructional video understanding. The competition example shows how our model can parse complex sequences of physical and emotional reactions, maintaining temporal coherence even in dynamic scenes. The eclipse footage example reveals VideoPASTA's ability to describe gradual visual transformations accurately, avoiding the baseline's tendency to oversimplify temporal transitions. Finally, the instruction scene identifying magic demonstrates our model's capability to establish clear causal relationships between actions and their outcomes, supported by specific visual evidence.

These qualitative results align with our quantitative findings, showing that VideoPASTA's structured approach to adversarial sampling leads to more precise and accurate video understanding across multiple dimensions. The improvements are especially evident in scenarios requiring temporal coherence, causal reasoning, and the integration of information across extended sequences. The results validate that our adversarial generation approach produces highly targeted examples that specifically challenge the intended aspects of video understanding, creating a focused and efficient learning signal for the model during preference optimization.

G Prompt Templates

The effectiveness of VideoPASTA depends heavily on the careful design of prompts that elicit targeted behaviors from generative models. Our prompt approach focuses on creating a framework that enables the consistent generation of high-quality preference pairs. Rather than using generic prompts that could lead to superficial or inconsistent responses, we develop a hierarchical strategy with explicit constraints and clear objectives. Each template (Figures 9–12) serves a distinct purpose in our pipeline while sharing a common structure that ensures consistency. The spatial misalignment template emphasizes physical relations that remain constant within local temporal windows. The temporal incoherence template focuses on capturing dynamic changes while maintaining causality. The cross-frame disconnection template bridges distant temporal connections without losing local context. Finally, the preference data filtering template

acts as a quality control mechanism, ensuring that our generated pairs maintain sufficient contrast while avoiding trivial differences. A key novelty in our method is the explicit incorporation of failure modes into the prompt design itself. Rather than hoping that models will naturally generate useful adversarial examples, we directly encode common pitfalls and misunderstandings into our adversarial prompt variants. The templates are designed to be model-agnostic, allowing them to work with different foundation models while maintaining consistent output quality.

Spatial Misalignment Prompt

You have a single video input. We want to test the model's spatial reasoning according to the following guidelines:

1. **Aligned Query Generation:**
    - Leverage world principles for spatial reasoning to produce 10 queries covering:
        - Occlusion (e.g., "Which object is partially hidden behind another?")
        - Depth perception (e.g., "Which item appears closest to the camera?")
        - Relative positioning ("How many objects occupy the left vs. right third of the frame?")
        - Foreground-background distinctions
        - Overall frame layout (top vs. bottom edges, etc.)
2. **Adversarial Query Generation:**
    - For each query, create an adversarial version.
    - Here, the video will be undersampled at 1 fps.
    - The adversarial query should actively induce spatial errors.
    - Example prompts:
        - If the query is about occlusion, force the model to claim everything is fully visible
        - if the query is about depth, insist all objects are equidistant

Hence, generate:

- "Straightforward Spatial Questions": 10 questions (as if asked under the normal sampling scenario)
- "Adversarial Variants": 3 matching adversarial instructions (3 per query) that lead the model to produce misaligned/spatially flawed responses.

Figure 9: Prompt template for generating aligned and adversarial spatial queries.

Temporal Incoherence Prompt

You have a single video input. We want to test the model's temporal reasoning according to the following guidelines:

1. **Aligned Query Generation:**
    - Leverage world principles for temporal reasoning ability on long videos to produce 10 queries covering:
        - Event ordering (e.g., "Which major action occurs first, and which follows?")
        - Action boundaries (e.g., "Does the person finish one task before starting the next?")
        - Transition points (e.g., "When does the subject switch activities?")
        - Causality (e.g., "Is the second event a direct result of the first?")
        - Concurrent actions (e.g., "Are there any simultaneous events, and how do they overlap?")
2. **Adversarial Query Generation:**
    - For each query, create an adversarial version.
    - Here, the video will be undersampled to induce temporal confusion.
    - The adversarial query should actively misrepresent event order, action boundaries, or causal links.
    - Example prompts:
        - Claim all actions occur at once, ignoring clear time gaps.
        - Collapse multiple sequential events into a single continuous action.

Hence, generate:

- "Straightforward Temporal Questions": 10 questions (as if asked under dense sampling and normal temporal clarity)
- "Adversarial Variants": 3 matching adversarial instructions (3 per query) that lead the model to produce temporally flawed or misaligned responses.

Figure 10: Prompt template for generating aligned and adversarial temporal queries.

Cross-Frame Disconnection Prompt

You have a single video input. We want to test the model's cross-frame (long-range) understanding according to the following guidelines:

1. **Aligned Query Generation:**
    - Please produce 10 queries covering:
        - Object continuity (e.g., does the same object appear in the opening and closing scenes?)
        - Character persistence (e.g., which participants return in later segments, and are they consistent with earlier roles?)
        - Setting evolution (e.g., does the location or environment change over time?)
        - Repeated actions (e.g., are certain actions performed in distant parts of the video, creating a parallel?)
        - Foreshadowing (e.g., do early events hint at outcomes shown near the end?)
2. **Adversarial Query Generation:**
    - For each query, create an adversarial version.
    - Deliberately break cross-frame connections by forcing the model to ignore continuity or treat identical objects/characters as unrelated.
    - Example prompts:
        - Insist that objects recurring in different scenes are completely different
        - Claim that characters present at both the start and end have no connection

Hence, generate:

- "Straightforward Cross-Frame Questions": 10 questions (as if the model respects full continuity across frames)
- "Adversarial Variants": 3 matching adversarial instructions (3 per query) that lead the model to produce disjointed or inconsistent responses across frames.

Figure 11: Prompt template for generating aligned and adversarial queries focusing on cross-frame video understanding.

Preference Data Filtering Prompt

You have a single video input and a set of four responses for each query:

1. One **aligned** response that is claimed to be well-aligned with the video content.
2. Three **adversarial** responses, each intentionally introducing spatial, temporal, or cross-frame errors.

The goal is to validate that:

- The *aligned* response truly aligns with the query (no unintended contradictions or inaccuracies).
- Each *adversarial* response introduces a clear misalignment without merely restating or slightly rephrasing the aligned response.

For each query and its four responses:

1. **Sanity-check the aligned response.**
   - Confirm that it accurately reflects the video's content in relation to the query.
   - If any errors or contradictions are detected, discard them.
2. **Examine each adversarial response.**
   - Identify whether it *deliberately* contradicts or distorts the query/video content (e.g., reversed sequence, false spatial claims).
   - If it is too similar to the aligned response or fails to demonstrate a clear misalignment, discard it.

Figure 12: Prompt template for validating one aligned and three adversarial responses to ensure robust preference pairs.

Adversarial QA Generation Prompt

You are tasked with generating adversarial video question-answering examples to test video language models' robustness. Based on the provided video, create:

1. **Adversarial Questions:**
    - Generate exactly 1 question per failure mode that cannot be reasonably answered from the video content.
    - These should appear legitimate but contain logical impossibilities or request information that is explicitly not present.
    - Target the following specific failure modes:
        (a) *Spatial Misalignment*: Request object relationships that don't exist (e.g., "How many people are standing behind the blue car?" when no blue car exists).
        (b) *Temporal Incoherence*: Ask about event sequences that violate the timeline (e.g., "What happens after the person leaves the room?" when no one leaves).
        (c) *Cross-Frame Disconnection*: Request connections between unrelated frames (e.g., "How does the opening scene connect to the dancing sequence?" when no dancing occurs).
2. **Adversarial Options:**
    - Create exactly 1 multiple-choice question per failure mode where all provided options are incorrect.
    - Questions should appear legitimate but all options should be misleading.
    - Include 4 plausible but incorrect options for each question.
    - The correct answer should always be "None of the Above" (not included in the options).
    - Target the same three failure modes as above.

Format each output as:

- **Adversarial Question [Spatial Misalignment]:** [Question text].
- **Adversarial Question [Temporal Incoherence]:** [Question text].
- **Adversarial Question [Cross-Frame Disconnection]:** [Question text].
- **Adversarial Options [Spatial Misalignment]:**
  Question: [Question text].
  Options:
  A. [Option 1].
  B. [Option 2].
  C. [Option 3].
  D. [Option 4].
  Correct Answer: None of the Above.
- **Adversarial Options [Temporal Incoherence]:**
  Question: [Question text].
  Options:
  A. [Option 1].
  B. [Option 2].
  C. [Option 3].
  D. [Option 4].
  Correct Answer: None of the Above.
- **Adversarial Options [Cross-Frame Disconnection]:**
  Question: [Question text].
  Options:
  A. [Option 1].
  B. [Option 2].
  C. [Option 3].
  D. [Option 4].
  Correct Answer: None of the Above.

Figure 13: Prompt template for generating adversarial questions and options across different failure modes.

Adversarial Question Evaluation Prompt

You are an expert evaluator for video question-answering systems. Your task is to determine whether a model correctly identifies unanswerable questions about a video.
Input:

1. Video context (description provided).
2. Adversarial question (deliberately unanswerable).
3. Model response to the question.

Evaluation criteria:

1. A correct response must identify that the question cannot be answered from the video.
2. Look for specific rejection patterns:
   (a) Explicit rejection phrases: "cannot be answered," "not possible to determine," "not shown in the video."
   (b) Identification of missing elements: "no such [object/action] appears," "insufficient information."
   (c) Recognition of logical impossibility: "the premise of the question is incorrect," "this scenario does not occur."
3. Incorrect responses attempt to answer the question by:
   (a) Hallucinating non-existent objects, actions, or relationships.
   (b) Making definitive statements about information not present in the video.
   (c) Failing to identify the adversarial nature of the question.

Output format:

1. **Judgment**: [CORRECT/INCORRECT].
2. **Reasoning**: Brief justification for your evaluation (1-2 sentences).
3. **Rejection Keywords Identified**: List specific rejection phrases used by the model.

Provide a binary decision (CORRECT/INCORRECT) based strictly on whether the model appropriately identified the question as unanswerable.

Figure 14: Prompt template for evaluating model responses to adversarial questions.

Adversarial Example Evaluation Prompt

Task: Evaluate whether the provided adversarial example correctly targets its intended failure mode in video understanding.
Query: [Original question asked about the video]
Aligned Response: [The correct/preferred response to the query]
Adversarial Example: [The adversarial example to be evaluated]
Claimed Failure Mode: [One of: "Spatial Misalignment", "Temporal Incoherence", or "Cross-Frame Disconnection"]
Failure Mode Definitions:

- **Spatial Misalignment**: Incorrectly describing spatial relations, object positions, occlusion patterns, depth, or relative positioning within a single frame.
- **Temporal Incoherence**: Violating the natural ordering of events, describing sequential actions as simultaneous, merging distinct events, or misordering the sequence of activities shown in the video.
- **Cross-Frame Disconnection**: Breaking object persistence across frames, describing the same object as different entities across scenes, failing to maintain character/object consistency, or incorrectly describing changes between distant frames.

Evaluation Instructions:

1. Carefully analyze the adversarial example in relation to the aligned response.
2. Determine if the adversarial example genuinely induces the claimed failure mode.
3. Your evaluation should be based solely on the definitions provided above.
4. Provide a binary judgment: "Yes" if the adversarial example correctly targets the claimed failure mode, "No" if it does not.
5. Briefly explain your reasoning (2-3 sentences).

Output Format:

```
Judgment: [Yes/No]
Reasoning: [Your brief explanation]
```

Figure 15: Prompt used for evaluating whether adversarial examples correctly target their claimed failure modes.