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 generated adversarial examples that deliberately violate spatial, temporal, or cross-frame relations. By applying Direct Preference Optimization to just 7,020 preference pairs, VideoPASTA learns robust representations that capture fine-grained spatial relationships and long-range temporal dynamics. Experiments on standard video benchmarks show significant relative performance gains of 3.05% on VideoMME, 1.97% on NeXTQA, and 1.31% on LongVideoBench, over the baseline Qwen2.5-VL model. 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 human annotation or captioning, relying on just 32-frame sampling, compared to the 96-frame, multi-GPU setups of prior work. This efficiency makes our approach a scalable, plug-and-play solution that seamlessly integrates with existing models while preserving their capabilities.*

1. Introduction

Recent advances in video large language models (Video-LLMs) have enabled efficient understanding and reasoning about videos [3, 7, 26, 30, 54, 55, 65]. These models capture both fine-grained details and broader contextual information in videos, achieving high performance in tasks such as captioning and question answering [6, 32, 41]. However, these models typically require large, high-quality annotated datasets and substantial computational resources for training. Recent work has explored instruction tuning as a strategy to reduce data dependency [7, 30, 31, 52, 73]. Instruction tuning refers to fine-tuning a pre-trained model on a curated dataset of instruction-response pairs, which allows the model to follow human directives. While recent efforts have successfully generated large instruction datasets through models like GPT-4V [73], the improvements in Video-LLMs achieved through training on these datasets have been limited. These models often show persistent failure modes in understanding and reasoning about videos, including spatial misalignment [9], temporal incoherence [9], and cross-frame disconnections [15, 20, 24, 35]. Addressing these alignment challenges through human annotation is prohibitively expensive, as it requires annotators to carefully identify and label examples that demonstrate proper grounding, temporal coherence, and factual consistency. This suggests that merely scaling up models and data may not be the most effective approach. Instead, the key challenge lies in ensuring faithful alignment between model responses and video content.

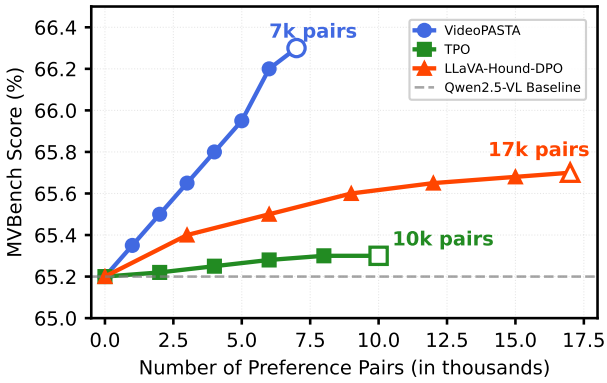


Figure 1. **With just 7k preference pairs**, VideoPASTA surpasses the Qwen2.5-VL [3] baseline, LLaVA-Hound [70] and TPO [28] on MVBench, demonstrating that targeted adversarial alignment can outperform larger but less focused datasets.

Recent work has explored video-language alignment [1, 28, 70] using Direct Preference Optimization (DPO) [39]. However, current approaches focus on collecting more preference data rather than addressing the core weaknesses

of Video-LLMs, generating data that reinforce existing strengths instead of challenging limitations. Furthermore, these methods rely on proprietary models for generating samples [70] or require video captioning [28], limiting their scalability.

This raises an important research question: *How can we align Video-LLMs to understand spatial, temporal, and cross-frame relationships without relying on human annotations, captions, or proprietary models while ensuring computational efficiency?* To address this challenge, we introduce **VideoPASTA** (**P**reference **A**lignment with **S**patio-**T**emporal-Cross Frame **A**dversaries), a novel framework that enforces accurate video-language alignment through targeted preference pairs. Our approach contrasts preferred responses with carefully generated adversarial responses that capture three common failure modes in video understanding: (1) **spatial misalignment**, where responses misrepresent object relationships and interactions within local frames, (2) **temporal incoherence**, where responses violate the natural progression of events, and (3) **cross-frame disconnection**, where responses incorrectly link information across distant frames. VideoPASTA leverages DPO and structured preference pairs to address failure modes in video understanding. In summary, our contributions are as follows:

1. We present VideoPASTA, a novel DPO-based framework for training video-language models, which addresses three key issues: spatial misalignment, temporal incoherence, and cross-frame disconnection while eliminating the need for human annotation, caption data, or proprietary models.
2. We set a new efficiency standard for preference optimization in video-language models, achieving substantial improvements with just 7,020 preference pairs compared to much larger instruction tuning datasets (1.3M) and preference datasets (17k) used in prior work.
3. We demonstrate through extensive evaluation across eight video understanding benchmarks that our targeted alignment approach leads to significant relative improvements, including +3.05% on VideoMME, +1.97% on NeXTQA, +1.69% on MVBench, and +1.31% on LongVideoBench compared to baselines.

2. Related Work

2.1. Video-LLMs

Recent advancements have led to the development of numerous Video-LLMs with impressive capabilities [7, 25, 26, 29, 42, 44, 51, 52, 55, 63, 68, 69, 73]. However, comprehensive evaluations [13, 33, 76] reveal persistent challenges in three critical areas. First, temporal reasoning remains a significant hurdle, particularly for long videos. Researchers have attempted to address this through increased context length [32, 42, 43, 69], compression techniques [22, 23, 29, 34, 53], and training-free approaches [18, 59, 62]. While these approaches improve token efficiency, they often fall short in enhancing fundamental temporal understanding. More specialized methods [6, 16, 41, 50, 57] directly target temporal reasoning but require substantial computational resources. Second, spatial misalignment remains a frequent challenge, as models struggle with object localization and spatial relationships, leading to incorrect positioning and poor occlusion handling [4, 40]. Third, cross-frame disconnection often results in failures in maintaining object continuity and narrative coherence across video segments [19, 48]. Most existing methods primarily address one of these dimensions or rely heavily on large-scale instruction-tuning datasets, which fail to resolve the core alignment challenges [5, 30, 51, 52, 71, 73].

VideoPASTA addresses all three failure modes through DPO-based training on structured preference pairs. Rather than addressing isolated weaknesses, we generate preference pairs that challenge the model's understanding across temporal, spatial, and cross-frame dimensions. This approach enables VideoPASTA to achieve more comprehensive video-language alignment compared to methods that rely on conventional instruction tuning.

2.2. Reward Modeling for Video-LLMs

Reward modeling is essential for aligning VLMs with human preferences by grounding their responses in an accurate visual context. Recent work has proposed a factually augmented RLHF approach that integrates image captions and verified data into the reward function to reduce hallucinations and improve output reliability [47]. Other approaches exploit AI-generated feedback via reinforcement learning to guide video–text alignment, thereby reducing reliance on expensive human annotations [2]. Complementary strategies in reward modeling for VLMs include the conditional preference optimization framework in MDPO [49], the calibrated self-rewarding approach that integrates visual constraints [77], and fine-grained reward modeling via dense sentence-level feedback in ViGoR [60]. In contrast, we employ DPO [39] to directly optimize preference learning, an approach adopted by recent Video-LLMs such as LLaVA-Hound-DPO [70] and i-SRT [1], which relies on proprietary models to generate samples and requires 17k preference pairs.

While previous approaches focus primarily on accumulating diverse preference samples, VideoPASTA fundamentally reimagines preference data generation through systematic targeting of failure modes. By generating preference pairs that deliberately challenge failure modes, we shift towards a preference quality over quantity approach. This reduces the required preference pairs and creates a more robust learning signal that simultaneously improves video understanding across all three critical dimensions.

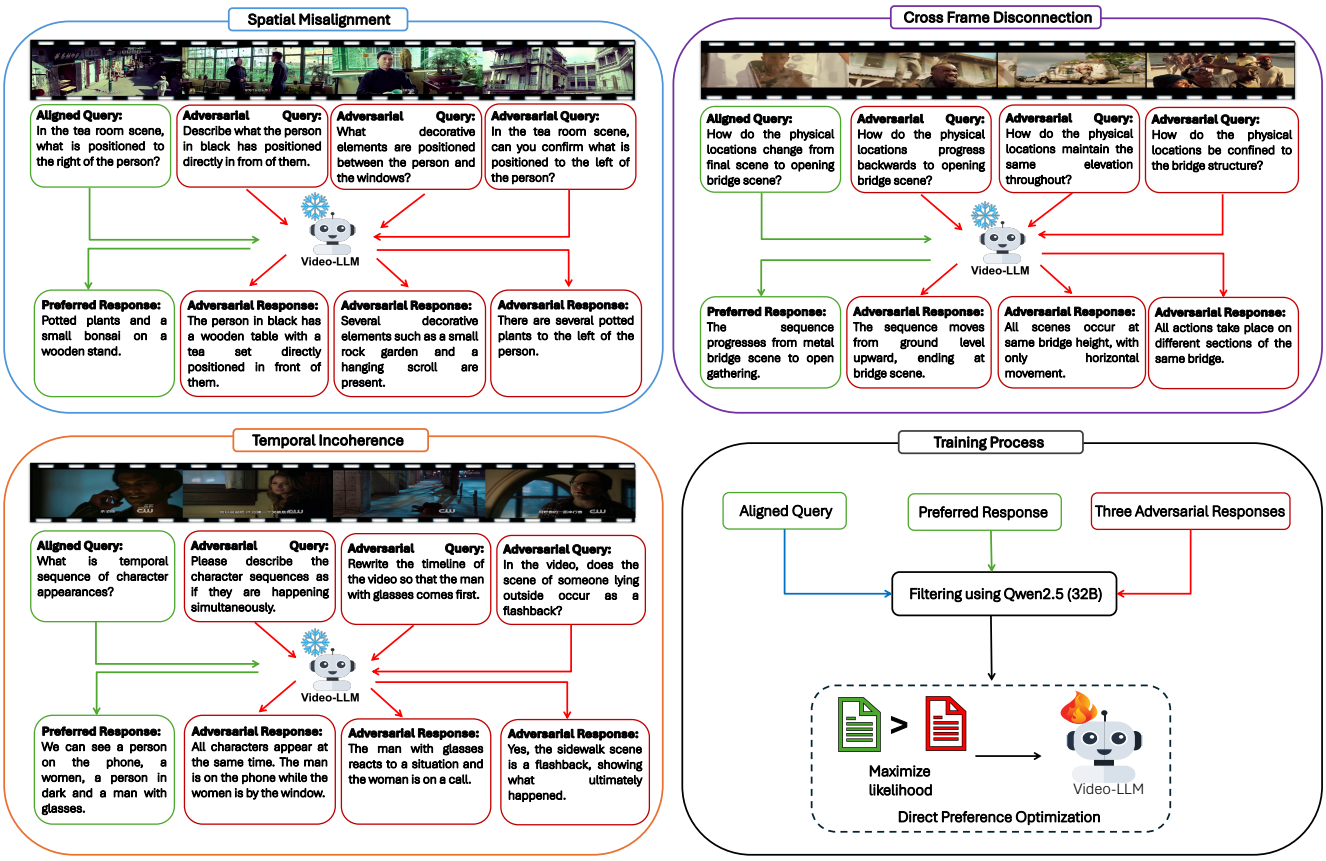


Figure 2. **Overview of VideoPASTA**. For each aligned query-response pair, we generate three types of targeted adversarial examples: (1) **Spatial Mislignment** that deliberately confuse object positions and relations (e.g., misplacing the plants relative to the person), (2) **Temporal Incoherence** that violate event sequences (e.g., claiming simultaneous actions that occur sequentially), and (3) **Cross-Frame Disconnection** that make invalid connections across distant frames (e.g., incorrectly describing location changes). These pairs are filtered using Qwen2.5-32B [61] to ensure quality, then used to train the model through DPO, maximizing the likelihood gap between preferred and adversarial responses. 🔥 → trainable, ❄ → frozen.

2.3. Self-Training

To address the challenge of large, annotated datasets, self-training methods [12, 21, 45, 46, 78] have been proposed to use unlabeled or weakly annotated data to improve model performance. However, current methods frequently depend on expensive GPT APIs [45] or require ground truth feedback [78], which limits scalability and hinders comprehensive alignment. While LLaVA-Hound-DPO [70] introduced preference optimization for video understanding, it operates primarily at the text level without directly targeting visual alignment failures and requires 17k preference pairs generated using proprietary models. Similarly, Temporal Preference Optimization (TPO) [28] improves temporal grounding by contrasting predictions from full videos with those from subsampled segments but addresses only temporal aspects while neglecting spatial relationships and cross-frame reasoning. Moreover, TPO still requires video captioning as an intermediate step and uses up to 10k preference pairs, introducing additional computational overhead.

What is missing from current approaches is a unified framework that efficiently targets all three critical dimensions of video understanding without requiring proprietary models, large preference datasets, or intermediate captioning steps. VideoPASTA fills this gap by employing preference pairs that challenge and refine a model's understanding, achieving superior alignment with just 7k carefully designed preference pairs.

3. VideoPASTA

VideoPASTA is a DPO-based framework that helps video-language models stay closely aligned with video content using structured preference optimization. Formally, our approach leverages a preference dataset D defined as follows:

$$D = \{(V, q, r^+, r^-)\}, \tag{1}$$

where V represents the input video, q denotes the query regarding the video content, r^+ is the preferred response accurately aligned with the video, and r^- is the adversarial response deliberately generated to introduce misalignment. Our approach addresses three key failure modes: spatial, temporal, and cross-frame misalignments by generating targeted adversarial examples for each. Specifically, we create one accurate ("aligned") example that correctly describes a scene, along with three misleading ("adversarial") examples, each designed to break a different aspect of alignment. By training on these contrastive pairs using DPO, the model learns to better distinguish between preferred responses and adversarial responses. As shown in Figure 2, this enables VideoPASTA to achieve more robust and comprehensive alignment across all aspects of video understanding. We focus on the following three failure modes of video understanding:

(1) Spatial Misalignment. Models often struggle with spatial relationships over short frame intervals. For example, a model may incorrectly describe a pedestrian as "crossing in front of" a moving vehicle when they are actually walking behind it, a critical spatial error that only becomes apparent through temporal context. Achieving accurate alignment requires understanding object interactions within local temporal contexts.

(2) Temporal Incoherence. Video frames capture dynamic events. In a cooking demonstration, a model might mistakenly describe a chef dicing vegetables, heating oil, and adding ingredients in a different order. To avoid temporal confusion, models must preserve the correct order of actions and their causal relationships.

(3) Cross-frame Disconnection. Distant frames provide context for continuity or setting changes. However, models often fail to link events, such as a character entering a building and later appearing in an office, treating them as unrelated scenes instead of a coherent progression. This results in incorrect generalizations and overlooked state changes, underscoring the need for robust cross-frame tracking.

To address each dimension, we generate adversarial responses that intentionally violate spatial, temporal, or cross-frame constraints. By contrasting these adversarial responses with preferred responses using DPO, the model is guided to learn precise alignment for each aspect of video-language understanding.

3.1. Spatial Misalignment

Spatial relationships, such as object positions, occlusions, depth order, and relative arrangements, are crucial for video understanding. To address these factors, we create targeted preference pairs that focus on spatial alignment.

Query Generation. We use InternVL2.5-38B [7] to automatically generate a variety of spatial queries (prompt provided in Appendix, Figure 10). These queries cover key aspects of spatial understanding, including occlusion (e.g., "Which object is partially hidden behind another?"), depth perception (e.g., "Which item appears closest to the camera?"), relative positioning (e.g., "How many objects occupy the left vs. right third of the frame?"), foreground-background relationships, and frame layout (e.g., "Which objects are at the top versus bottom edge?").

Preferred Response Generation. To capture fine-grained spatial details, we sample the input video at 32 fps. The baseline Qwen2.5-VL [3] model is then explicitly prompted to provide coherent and detailed descriptions of object interactions, depth cues, and spatial configurations across neighboring frames. This ensures that the generated preferred responses accurately reflect the true spatial relationships in the video.

Adversarial Response Generation. We generate challenging examples by undersampling videos at 1 fps and modifying prompts to induce spatial errors. For occlusion queries, adversarial prompts instruct the baseline model to describe all objects as fully visible, even when occluded (e.g., stating "the person is fully visible" when partially hidden behind a counter). For depth-related queries, prompts require the model to reason that all objects are equidistant, disregarding clear perspective cues (e.g., claiming "the cup and the mountain are at the same distance from the camera"). These intentionally flawed responses act as adversaries to test and refine precise spatial reasoning.

3.2. Temporal Incoherence

Accurately capturing event sequences, action transitions, and causal relationships is crucial for robust video understanding. To evaluate and enhance temporal reasoning, we create preference pairs that focus on the model's temporal coherence.

Query Generation. We use InternVL2.5-38B to generate queries that focus on key aspects of temporal understanding (prompt provided in Appendix, Figure 11). These include event ordering (e.g.,"Which major action occurs first, and which follows?"), action boundaries (e.g., "Does the person complete 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?"), and concurrent actions (e.g., "Are there any simultaneous events, and how do they overlap?").

Preferred Response Generation. To obtain temporally coherent responses, we process the video at its native frame rate, ensuring dense temporal sampling. The baseline model is prompted to describe the precise sequence of events, effectively capturing transitions, dependencies, and causal relationships.

Adversarial Response Generation. We create adversarial examples by undersampling the videos at 1 fps and utilizing prompts that induce temporal distortions. One adversarial prompt instructs the baseline model to describe sequential actions as occurring simultaneously (e.g., "the chef chopping vegetables, heating the pan, and plating the dish all at the same time"). Another prompts the model to ignore action transitions, merging distinct events into a continuous sequence (e.g., describing "a continuous swimming motion" when the video shows separate diving, swimming, and exiting phases). These flawed responses act as adversaries to test and refine precise temporal reasoning.

3.3. Cross-frame Disconnection

Robust video understanding requires capturing long-range cross-frame relationships, such as object continuity, character persistence, setting changes, and narrative progression across distant frames. To achieve this, we create targeted preference pairs that focus specifically on these cross-frame dependencies.

Query Generation. We use InternVL2.5-38B to automatically generate queries that evaluate various aspects of cross-frame understanding (prompt provided in Appendix, Figure 12). For instance, queries ask whether the same object appears in both the opening and closing scenes, whether characters present early in the video reappear, whether the setting evolves over time, whether actions repeat across distant segments, and how early events foreshadow later developments.

Preferred Response Generation. To capture coherent long-range dependencies, we uniformly sample the entire video sequence at its native frame rate. The baseline model is prompted to describe consistent object transformations, character developments, setting changes, and narrative connections, ensuring that preferred responses accurately reflect the continuity and narrative progression throughout the video.

Adversarial Response Generation. We generate challenging examples by undersampling the video at 1 fps by using prompts that disrupt cross-frame connections. For object continuity, adversarial prompts instruct the baseline model to treat identical objects in different scenes as unrelated (e.g., describing "a new red car appears" when it is the same vehicle shown from a different angle). For character persistence, prompts require the model to treat similar-looking characters as entirely distinct (e.g., stating "a different person enters the office" when it is clearly the same individual from earlier). These flawed responses act as adversaries to test and strengthen robust, long-range video-language alignment.

3.4. Preference Data Filtering

We generate three adversarial examples for each failure mode (spatial, temporal, cross-frame) per query and use Qwen2.5-32B [61] 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 (prompt provided in Appendix, Figure 13). Adversaries that are too similar to the aligned samples or lack clear contradictions are discarded and regenerated. Similarly, we perform a "sanity check" on preferred 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. Each data point in the preference dataset is a tuple (V, q, r^+, r^-), where V represents the input video, q denotes the query regarding the video content, r^+ is the preferred response, and r^- is the adversarial response.

For a video-language model M_θ, we define the DPO loss for a single preference pair as:

$$\begin{aligned}\Delta(V, q, r^+, r^-) &= \log p_\theta(r^+ \mid V, q) - \log p_\theta(r^- \mid V, q),\\ \mathcal{L}_{\text{DPO}}(V, q, r^+, r^-) &= -\log \sigma\Big(\lambda\, \Delta(V, q, r^+, r^-)\Big),\end{aligned} \tag{2}$$

where σ is the sigmoid function, λ is a scaling factor, and the term inside the brackets represents the log probability difference between the preferred and adversarial responses. We then compute the DPO loss separately for each subset of preference pairs and weight them by α, β, and γ to reflect the relative importance of spatial, temporal, and cross-frame alignment. The overall training objective is:

$$\mathcal{L} = \alpha\, \mathbb{E}_{\mathcal{D}_s}\big[\mathcal{L}_{\text{DPO}}\big] + \beta\, \mathbb{E}_{\mathcal{D}_t}\big[\mathcal{L}_{\text{DPO}}\big] + \gamma\, \mathbb{E}_{\mathcal{D}_c}\big[\mathcal{L}_{\text{DPO}}\big]. \tag{3}$$

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 [64]. Our structured adversarial sampling pipeline produces $90,000$ preference pairs which after filtering are reduced to $7,020$ (additional details included in Appendix §A.3). We fine-tune Qwen2.5-VL [3] with the SWIFT [75] framework for efficient adaptation. Training and evaluation are performed on two NVIDIA L40S GPUs (48GB), with a maximum frame limit of 32 to avoid CUDA OOM errors.

Model	TempCompass	PerceptionTest	NeXTQA	MVBench	Vinoground	MLVU	LongVideoBench	VideoMME
(1) Baseline Models								
Qwen2.5-VL [3]	71.7	68.6	75.8	65.2	27.9	68.7	60.7	62.2
+ SFT	71.8	69.1	77.2	65.5	28.0	68.8	60.9	62.5
+ Hound-DPO [70]	70.3	67.6	76.1	65.7	27.8	66.4	56.3	63.2
+ TPO [28]	71.5	69.0	77.6	65.3	28.1	68.9	59.2	**64.2**
(2) State-of-the-Art Models								
VideoLLaMA2† [8]	43.4	51.4	-	54.6	22.1	35.5	-	47.9
Kangaroo† [32]	-	-	-	61.0	–	61.0	54.8	56.0
LLaVA-NeXT-Video† [71]	53.0	48.8	53.5	53.1	17.8	-	49.1	46.5
LongVA† [69]	-	-	-	-	–	58.8	51.3	52.6
LLaVA-NeXT-Interleave [26]	54.1	51.2	67.0	46.5	15.2	52.5	44.8	48.3
InternVL2.5 [7]	68.3	62.2	77.0	**69.8**	**29.4**	59.5	52.9	57.9
Qwen2-VL [52]	68.9	62.3	75.7	64.9	24.8	57.5	55.6	55.3
LLaVA-OneVision [25]	64.5	57.1	**79.3**	56.7	25.0	64.9	56.3	58.2
LLaVA-Video [73]	66.4	67.9	74.2	58.6	24.3	66.5	58.2	62.4
(3) Preference-Optimized Models								
LLaVA-Hound-DPO [70]	55.5	45.1	61.6	36.6	23.8	41.1	36.7	34.2
i-SRT [1]	56.0	47.0	63.0	36.3	23.7	39.9	38.2	34.7
LLaVA-Video-TPO [28]	66.6	66.3	77.8	56.7	24.0	66.3	58.3	62.4
VideoPASTA	**72.3**$_{+0.84\%}$	**69.4**$_{+1.17\%}$	77.3$_{+1.97\%}$	66.3$_{+1.69\%}$	28.3$_{+1.43\%}$	**69.2**$_{+0.73\%}$	**61.5**$_{+1.31\%}$	64.1$_{+3.05\%}$

Table 1. **Comprehensive evaluation of VideoPASTA against leading video understanding models**. We compare VideoPASTA with (1) baseline models built on Qwen2.5-VL, (2) current state-of-the-art models, and (3) models enhanced through preference optimization. The best scores are in **bold**, and the second-best scores are underlined. For VideoPASTA, relative performance improvements over the Qwen2.5-VL foundation model are indicated as subscripts. All results, except those marked with †, are reproduced using LMMs-Eval [67].

We use LoRA [17] with $\alpha = \beta = 8$ and a DPO scaling factor of $\beta = 0.1$. The model converges in one epoch, requiring just 6 hours of training. Evaluation is done using LMMs-Eval [67] for a fair comparison to prior work. Additional experiments on data quality (§A.2), preference learning efficiency (§A.2.2), dataset samples (§A.3.2), qualitative samples (§A.4) and prompt design (§A.5) are provided in appendix.

We evaluate on general video understanding benchmarks: TempCompass [33] (temporal understanding), PerceptionTest [37] (visual perception), NeXTQA [58] (compositional reasoning), MVBench [27] (multi-task reasoning), and Vinoground [66] (dense temporal reasoning). For long-form evaluation, we use LongVideoBench [56] (hour-long videos), MLVU [76] (multi-task, 3-minute to 2-hour videos), and VideoMME [13] (6 visual domains, 30 subfields, 11-second to 1 hour videos).

4.1. Results

We compare VideoPASTA with (1) baseline models built on Qwen2.5-VL [3], (2) current state-of-the-art models, and (3) models enhanced through preference optimization. Table 1 presents the evaluation results.

Gains Over Foundation Models. Compared to its backbone model, Qwen2.5-VL, VideoPASTA achieves significant relative improvements on key benchmarks: VideoMME (+3.05%), Vinoground (+1.43%), PerceptionTest (+1.17%), and LongVideoBench (+1.31%). These gains reflect enhanced temporal reasoning, fine-grained visual understanding, and improved temporal coherence in long-form video analysis. For baseline models built on Qwen2.5-VL, we first evaluate supervised fine-tuning (SFT) using aligned examples (i.e., queries and preferred responses) but observe only marginal improvements. This indicates that aligned examples alone are insufficient, and models need to learn from mistakes through adversarial examples. Hound-DPO [70] uses proprietary models to generate samples and requires 17k pairs, while TPO [28] focuses only on temporal alignment with 10k pairs and needs captioning data. In comparison, VideoPASTA performs better with just 7k pairs, showing that carefully chosen adversarial examples can achieve more effective alignment than larger, less focused datasets.

Comparison with State-of-the-Art. VideoPASTA outperforms state-of-the-art models on four of eight video understanding benchmarks. Key improvements include a 5.85% relative gain over InternVL2.5 in temporal reasoning on TempCompass (72.3%) and a 2.2% increase on PerceptionTest (69.4%) over LLaVA-Video, which was instruction-tuned on 1.3M SFT pairs. Our method also shows significant improvements on long-form video benchmarks, including LongVideoBench (61.5%, +5.67% vs. LLaVA-Video), MLVU (69.2%, +4.06%), and VideoMME (64.1%, +2.72%).

Outperforming Preference-Optimized Models. Compared to preference-optimized models, VideoPASTA outperforms LLaVA-Hound-DPO [70] and i-SRT [1] on all

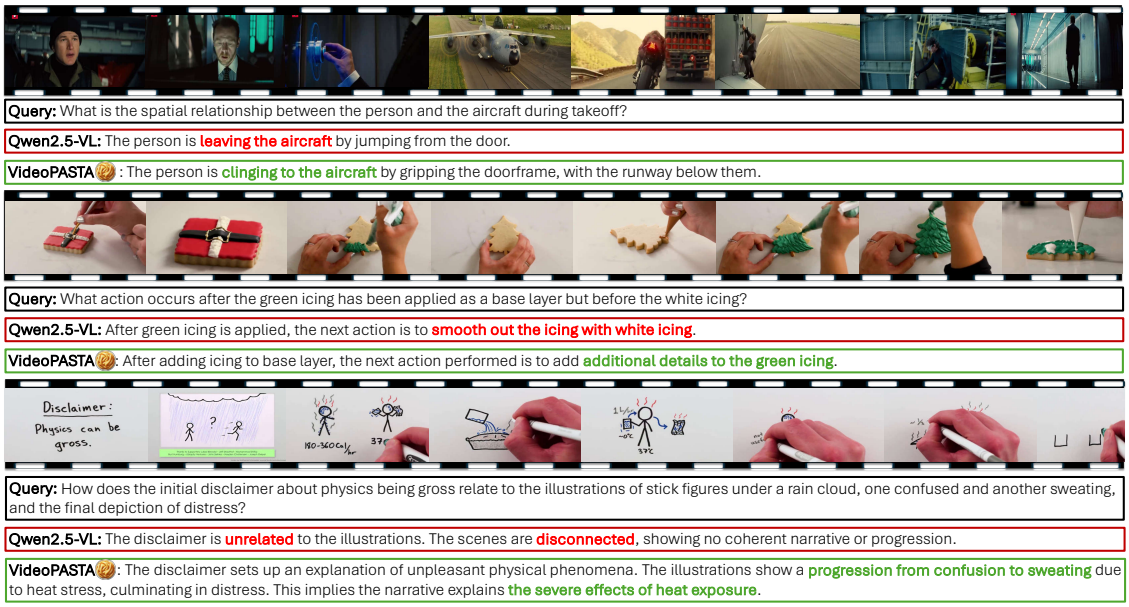


Figure 3. **Qualitative comparison of VideoPASTA against baseline Qwen2.5-VL** [3]. The examples demonstrate how our method improves different aspects of video understanding: (1) *Spatial reasoning*: Correctly identifies the spatial relationship between person and aircraft (clinging to doorframe vs. incorrectly stating "leaving"), (2) *Temporal understanding*: Accurately captures the sequence of decorative actions (adding details to base layer vs. incorrectly jumping to white icing), (3) *Cross-frame reasoning*: Successfully connects thematic progression across frames (heat stress narrative) while the baseline model fails to establish relationships between scenes.

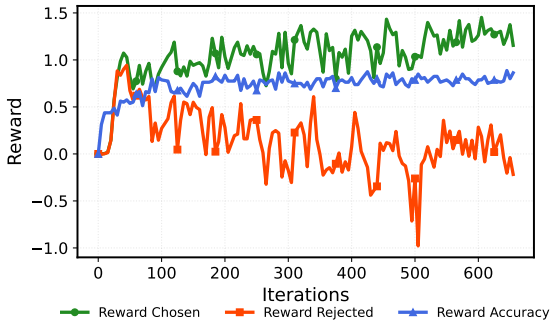


Figure 4. DPO training converges on well-grounded responses, with **chosen rewards rising**, **rejected rewards falling**, and overall reward accuracy stabilizing over iterations.

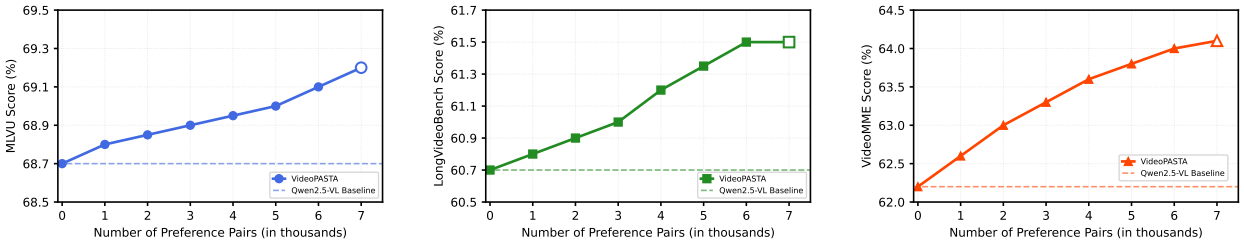
eight benchmarks and outperforms LLaVA-Video-TPO [28] on seven out of eight benchmarks by a significant margin. These improvements highlight that our multi-dimensional approach, which tackles critical failure modes in Video-LLMs, addresses fundamental challenges in prior work while requiring minimal resources.

4.2. Ablation Studies

DPO Training Dynamics. To illustrate how our DPO-based preference optimization evolves over time, we plot key reward metrics in Figure 4. Specifically, we track (1) *reward accuracy*, which measures how often the model ranks a preferred response higher than its adversarial counterpart, (2) *reward for chosen responses*, and (3) *reward for rejected responses*. The model quickly learns to distinguish correct from incorrect interpretations, as indicated by an early rise in reward accuracy. Over successive iterations, accuracy stabilizes around 70–75%, suggesting that the model consistently favors preferred responses over adversarial ones. Meanwhile, the increasing gap between chosen and rejected rewards confirms that structured adversarial sampling reinforces clear preference boundaries. This behavior highlights the effectiveness of our DPO-based approach in aligning the model with desired video-language interpretations.

Efficiency of Preference Data Scaling. Figure 5 shows how performance scales with the amount of preference data. Even with just 1,400 pairs, VideoPASTA surpasses Qwen2.5-VL on MLVU, VideoMME, and LongVideoBench. Performance improves steadily across all benchmarks as more preference pairs are added, with noticeable gains observed between 3k and 4k pairs, and the upward trend continuing through 7k pairs. This steady improvement across diverse benchmarks indicates that structured adversarial sampling effectively improves video understanding without requiring excessive amounts of preference data.

Effectiveness of Targeted Adversarial Sampling. Table 2 demonstrates that each failure mode in our adversarial sampling pipeline contributes distinct advantages. Training ex-



(a) MLVU (b) LongVideoBench (c) VideoMME

Figure 5. **Impact of preference data size on model performance**. We evaluate our model's performance across three benchmarks while varying the number of preference pairs used for training.

Model	Temporal	Spatial	Action	Object
Qwen2.5-VL [3]	35.0	63.6	54.0	55.9
w/ Temporal Only	44.0	67.1	49.2	51.3
w/ Spatial Only	40.1	74.8	55.0	55.4
w/ Cross-Frame Only	43.3	66.8	54.9	57.2
Full Model	**45.2**$_{+29.14\%}$	**78.6**$_{+23.58\%}$	**56.1**$_{+3.89\%}$	**58.4**$_{+4.47\%}$

Table 2. **Effect of each targeted failure mode on VideoMME**. Subscripts show percentage increases from Qwen2.5-VL.

Model	MLVU	LongVideoBench	VideoMME
VideoLLaVA [30]	47.3	39.1	39.9
+ VideoPASTA	**50.2**$_{+6.13\%}$	**42.4**$_{+8.44\%}$	**40.5**$_{+1.50\%}$
LLaVA-NeXT-Interleave [26]	52.5	44.8	48.3
+ VideoPASTA	**54.0**$_{+2.85\%}$	**45.5**$_{+1.56\%}$	**49.9**$_{+3.31\%}$
LLaVA-OneVision [25]	64.9	56.3	58.2
+ VideoPASTA	**66.3**$_{+2.16\%}$	**56.8**$_{+0.89\%}$	**60.9**$_{+4.64\%}$

Table 3. **Model-agnostic improvements from VideoPASTA**.

clusively with temporal samples significantly improves temporal reasoning (+25.71%) but minimally impacts spatial or object reasoning. Similarly, spatial-only training boosts spatial reasoning the most (+17.61%), while cross-frame samples enhance both object (+2.32%) and temporal (+23.71%) reasoning. Combining all three modes delivers the best overall performance, with significant improvements in temporal (+29.14%), spatial (+23.58%), action (+3.89%), and object (+4.47%) reasoning. These results confirm that adversarial sampling across multiple failure modes provides complementary benefits, leading to more holistic video-language alignment than focusing on each mode individually.

Model-Agnostic Improvements. Table 3 shows that VideoPASTA consistently improves performance across different base models. We apply our framework to VideoLLaVA [30], LLaVA-NeXT-Interleave [26], and LLaVA-OneVision [25]. In each case, VideoPASTA enhances benchmark scores on MLVU, LongVideoBench, and VideoMME, confirming that our structured adversarial sampling strategy is effective regardless of the underlying model. These results highlight the robustness of VideoPASTA as a model-agnostic approach for improving video-language alignment.

Adversarial Examples per Aligned Sample	MLVU	LongVideoBench	VideoMME
1	67.5	58.8	62.3
2	68.4	60.2	63.2
3 (Ours)	**69.2**	**61.5**	**64.1**
4	68.9	61.2	63.8
5	68.7	61.0	63.6

Table 4. **Effect of the number of adversarial examples per aligned sample.** Performance peaks at 3 adversarial examples, matching our three targeted failure modes (spatial, temporal, and cross-frame misalignment).

Frame Sampling	MLVU	LongVideoBench	VideoMME
32:32	68.7	60.8	62.4
16:16	68.6	60.7	62.3
32:8	69.0	61.2	63.5
16:4	68.9	61.0	63.2
32:1 (Ours)	**69.2**	**61.5**	**64.1**
16:1	69.0	61.3	63.7

Table 5. **Effect of frame sampling rates.** Best results in **bold**.

Number of Adversarial Examples per Aligned Sample. Table 4 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 preferred response with exactly three adversarial responses, one for each failure mode, best reinforces alignment across spatial, temporal, and cross-frame reasoning.

Frame Sampling. Our analysis of sampling rates (Table 5) 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

Model	Spatial Misalignment		Temporal Incoherence		Cross-Frame Disconnection	
	Adv. Question (%)	Adv. Options (%)	Adv. Question (%)	Adv. Options (%)	Adv. Question (%)	Adv. Options (%)
Qwen2.5-VL [3]	38.4	42.6	35.2	39.5	31.8	36.7
LLaVA-Hound-DPO [70]	39.2	43.1	36.5	39.8	31.9	37.2
TPO [28]	41.3	44.5	48.2	51.4	32.4	37.5
VideoPASTA	**46.8**$_{+21.88\%}$	**51.1**$_{+19.95\%}$	**49.7**$_{+41.19\%}$	**52.8**$_{+33.67\%}$	**33.1**$_{+4.09\%}$	**38.2**$_{+4.09\%}$

Table 6. **Performance on Adversarial QA Samples Across Different Failure Modes**. Adv. Question: Adversarial questions where no correct answer exists (higher rejection rate is better). Adv. Options: Adversarial options where "None of the Above" is the correct answer (higher NOTA selection rate is better). Each cell shows the percentage of correctly handled adversarial samples. Subscripts show percentage increases from Qwen2.5-VL.

Model	MLVU	LongVideoBench	VideoMME
Qwen2-VL (2B)	51.3	46.6	50.1
+ **VideoPASTA**	51.9$_{+1.16\%}$	47.7$_{+2.36\%}$	51.2$_{+2.19\%}$
Qwen2-VL (7B)	**57.5**	**55.6**	**55.3**
Qwen2.5-VL (3B)	65.9	55.8	61.2
+ **VideoPASTA**	66.7$_{+1.21\%}$	56.0$_{+0.35\%}$	61.8$_{+0.98\%}$
Qwen2.5-VL (7B)	**68.7**	**60.7**	**62.2**

Table 7. **Preference Learning vs. Model Scaling**.

creates clear misalignment patterns. This result is consistent across benchmarks, highlighting the importance of a well-designed sampling strategy in model training.

Robustness Against Adversarial Inputs. To evaluate VideoPASTA's robustness against failure modes, we tested 100 videos from LLaVA-Video [72] using GPT-4o [36] (prompt provided in Appendix, Figure 14) to generate both adversarial questions (unanswerable queries) and adversarial options (where "None of the Above" is correct) per failure mode. As shown in Table 6, VideoPASTA significantly outperforms baselines across all categories, with the most substantial gains in temporal reasoning (+41.19%). 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 15), specifically identifying rejection phrases like "cannot be answered" and "insufficient information" when models correctly recognized adversarial inputs.

Preference Learning vs. Model Scaling. Table 7 presents the comparative analysis between preference learning and model scaling. VideoPASTA demonstrates consistent performance improvements across model sizes, with relative gains of 1.16-2.36% on smaller models (2B parameters) and 0.35-1.21% on larger models (3B parameters). This experiment provides two critical insights: (1) preference-based alignment offers an orthogonal optimization path to parameter scaling, and (2) our targeted adversarial sampling effectively improves alignment regardless of model capacity. Notably, applying VideoPASTA to Qwen2.5-VL (3B) yields performance comparable to vanilla Qwen2.5-VL (7B) on VideoMME, achieving 99.4% of the larger model's performance with only 42.9% of the parameters. These results validate that addressing core failure modes through structured preference optimization provides an efficient alternative to computational-intensive scaling approaches, particularly beneficial in resource-constrained environments.

5. Conclusion

We introduce VideoPASTA, a novel DPO-based framework that improves video-language models through structured adversarial sampling targeting spatial, temporal, and cross-frame misalignments. Our approach sets a new efficiency standard, achieving significant performance gains across eight benchmarks with just 7,020 preference pairs without the need for human supervision or video captioning. This model-agnostic method operates efficiently with 32-frame sampling, demonstrating that targeted adversarial examples enable more effective learning than generic instruction tuning. While individual failure modes lead to improvements in their respective dimensions, combining all three results in significant gains across all aspects of video understanding. Our method challenges the reliance on massive datasets and paves the way for resource-efficient video-language alignment. Future research could focus on evaluation metrics for aligned models, providing clearer insights into their reasoning process.

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

[1] Daechul Ahn, Yura Choi, San Kim, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. i-srt: Aligning large multimodal models for videos by iterative self-retrospective judgment. *arXiv preprint arXiv:2406.11280*, 2024. 1, 2, 6

[2] Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 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, 2024. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 4, 5, 6, 7, 8, 9, 15

[4] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, and et al. 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, 2024. 2

[5] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 2

[6] Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024. 1, 2

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 2, 4, 6

[8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 6

[9] Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. Vidhal: Benchmarking temporal hallucinations in vision llms. *arXiv preprint arXiv:2411.16771*, 2024. 1

[10] Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. *arXiv preprint arXiv:2410.14148*, 2024. 13

[11] Shijian Deng, Wentian Zhao, Yu-Jhe Li, Kun Wan, Daniel Miranda, Ajinkya Kale, and Yapeng Tian. Efficient self-improvement in multimodal large language models: A model-level judge-free approach. *arXiv preprint arXiv:2411.17760*, 2024. 13

[12] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*, 2024. 3

[13] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2, 6

[14] Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. Tldr: Token-level detective reward model for large vision language models. *arXiv preprint arXiv:2410.04734*, 2024. 13

[15] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2024. 1

[16] Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos. *arXiv preprint arXiv:2411.14901*, 2024. 2

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[18] Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding. *arXiv preprint arXiv:2502.06428*, 2025. 2

[19] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14271–14280, 2024. 2

[20] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2024. 1

[21] Yogesh Kulkarni and Pooyan Fazli. Videosavi: Self-aligned video language models without human supervision. *arXiv preprint arXiv:2412.00624*, 2024. 3

[22] Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Vidcompress: Memory-enhanced temporal compression for video understanding in large language models. *arXiv preprint arXiv:2410.11417*, 2024. 2

[23] Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, and Xinyu Li. Video token merging for long-form video understanding. *arXiv preprint arXiv:2410.23782*, 2024. 2

[24] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, et al. 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)*, 2024. 1

[25] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 6, 8

[26] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1, 2, 6, 8

[27] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, and et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206. IEEE, 2024. 6

[28] Rui Li, Xiaohan Wang, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*, 2025. 1, 2, 3, 6, 7, 9, 14

[29] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 2

[30] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pages 5971–5984. Association for Computational Linguistics, 2024. 1, 2, 8

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[32] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. 1, 2, 6

[33] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics (ACL)*, pages 8731–8772. Association for Computational Linguistics, 2024. 2, 6

[34] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024. 2

[35] Xinyu Ma, Yifan Li, Hao Wang, et al. 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)*, 2024. 1

[36] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 9

[37] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2024. 6

[38] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language models with bootstrapped preference optimization. In *Proceedings of ECCV 2024*, 2024. 13

[39] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

[40] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. 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, 2024. 2

[41] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 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, 2024. 1, 2

[42] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 2

[43] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 2

[44] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 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, 2024. 2

[45] Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. Stllava-med: Self-training large language and vision assistant for medical. *arXiv preprint arXiv:2406.19973*, 2024. 3

[46] Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. Sq-llava: Self-questioning for large vision-language assistant. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 156–172, 2025. 3

[47] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, and et al. Aligning Large Multimodal Models with Factually Augmented RLHF. In *Findings of the Association for Computational Linguistics (ACL)*, pages 13088–13110, 2024. 2

[48] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, and et al. Koala: Key frame-conditioned long video-llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13581–13591, 2024. 2

[49] Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 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. ACL, 2024. 2

[50] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024. 2

[51] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 2

[52] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 6

[53] Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*, 2024. 2

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

[55] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. In *ECCV*, 2024. 1, 2

[56] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 6

[57] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. *arXiv preprint arXiv:2411.10332*, 2024. 2

[58] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 6

[59] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024. 2

[60] Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *Proceedings of ECCV 2024*, 2024. 2

[61] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 3, 5, 15

[62] Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. Vca: Video curious agent for long video understanding. *arXiv preprint arXiv:2412.10471*, 2024. 2

[63] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2

[64] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 9127–9134, 2019. 5, 15

[65] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 543–553, 2023. 1

[66] Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024. 6

[67] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. 6

[68] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, and et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *CoRR*, abs/2407.03320, 2024. 2

[69] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2, 6

[70] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 1, 2, 3, 6, 9, 14

[71] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 2, 6

[72] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 9

[73] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 2, 6

[74] Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, and et al. Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*, 2024. 13

[75] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. 5

[76] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 2, 6

[77] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024. 2

[78] Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. Video-star: Self-training enables video instruction tuning with any supervision. *arXiv preprint arXiv:2407.06189*, 2024. 3

A. Appendix

A.1. Reward Modeling for VLMs

Several recent works have proposed alternative strategies for reward modeling in VLMs. Deng et al. [11] introduce a judge-free self-improvement framework that leverages controlled hallucination and lightweight contrastive encoders to generate high-quality preference pairs, reducing dependence on large-scale model judges. Cui et al. [10] recast reward modeling as a next-token prediction problem by exploiting fine-grained visual signals, which allows token-level reward feedback for more precise model alignment. Similarly, Fu et al. [14] develop a token-level detective reward model that provides detailed, interpretable rewards for each token, thereby facilitating effective self-correction and hallucination detection. Pi et al. [38] address pretraining biases in multimodal models by bootstrapping adversarial responses, using both image distortion and LLM bias injection, to suppress false correlations from the pretraining phase. Extending these ideas to the robotics domain, Zhang et al. [74] propose GRAPE, which aligns robot policies through trajectory-level reward modeling. Together, these approaches offer complementary insights into fine-grained, efficient, and scalable reward modeling for VLMs. In contrast, our work leverages structured adversarial sampling to generate synthetic preference data that specifically targets common failure modes in video-language alignment. This targeted strategy enables our model to refine spatial, temporal, and object-level representations more effectively, delivering robust multi-dimensional video understanding without the need for costly human annotations or reliance on external proprietary models.

A.2. Quality Analysis of Preference Data

A.2.1. Failure Mode Targeting Accuracy

To assess how well our adversarial examples target their intended failure modes, we conduct a systematic evaluation using GPT-4o as an independent judge. We randomly sample 200 preference pairs from our training dataset, each containing an aligned query-response pair and three corresponding adversarial examples (one targeting each failure mode). For each adversarial example, we prompt GPT-4o to determine whether it correctly induces the specific failure mode it is designed to target without revealing the intended category to prevent bias.

The results, presented in Table 8, demonstrate strong targeting precision across all three categories. Spatial misalignment examples achieve the highest targeting accuracy at 96.1%, indicating our approach excels at generating examples that specifically challenge spatial relationship understanding. Temporal incoherence examples show strong performance at 92.4%, while cross-frame disconnection examples reach 88.3% accuracy. The slightly lower performance in cross-frame cases likely reflects the inherent complexity of maintaining coherent object continuity across distant frames. The prompt is provided in Figure 16.

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

Table 8. Failure Mode Targeting Accuracy by Category

A.2.2. Preference Optimization Efficiency

Figures 6 and 7 present a comprehensive analysis of preference learning efficiency across methods, revealing several key insights into the relationship between data quantity, quality, and performance gains.

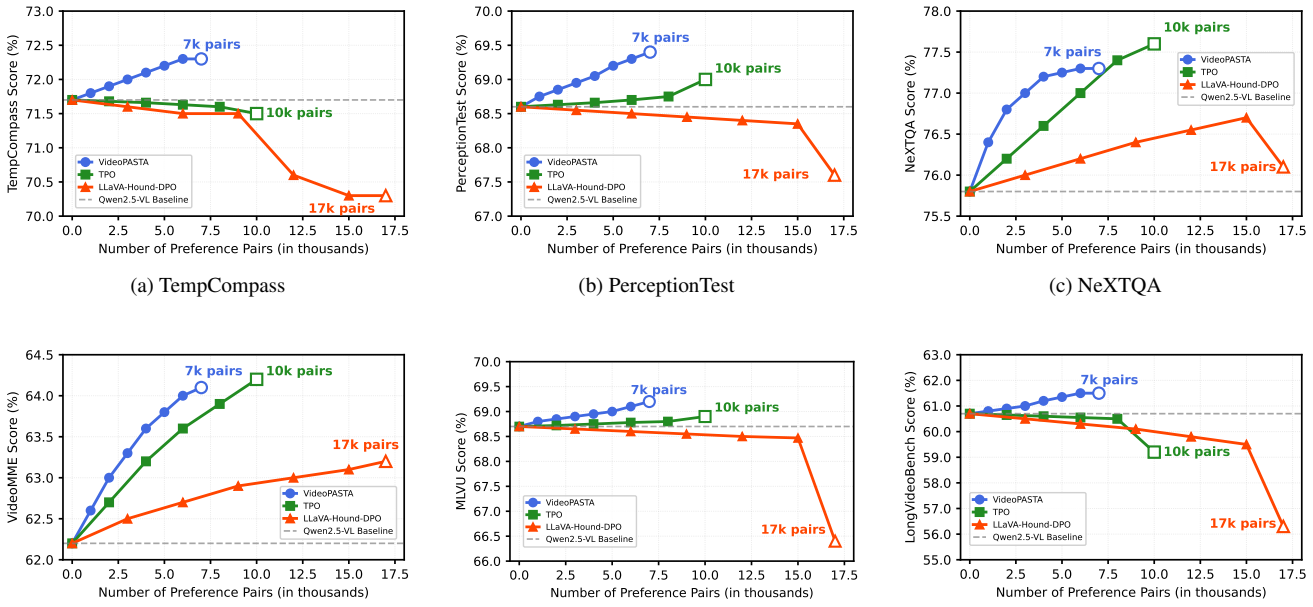
Consistent Efficiency Across Benchmarks. Across all benchmarks, VideoPASTA consistently demonstrates superior data efficiency, achieving comparable or better performance than alternative methods while using significantly fewer preference pairs. This efficiency is particularly evident in temporal understanding benchmarks like TempCompass, where VideoPASTA reaches 72.3% with just 7k pairs, outperforming TPO (71.5% with 10k pairs) by a substantial margin. Similarly, on spatial understanding benchmarks like PerceptionTest, VideoPASTA achieves 69.4% with 7k pairs, effectively matching TPO's 69.0% performance at 10k pairs.

Information Gain Quantification. To precisely measure learning efficiency, we calculate the information gain $\mathcal{G}$, defined as:

$$\mathcal{G}_{\text{model}} = \frac{S_{\text{final}} - S_{\text{baseline}}}{n/1000} \tag{4}$$

where S_{final} is the model's performance score after training, S_{baseline} is the baseline performance, and n is the number of preference pairs used. This metric quantifies performance improvement per thousand training examples.

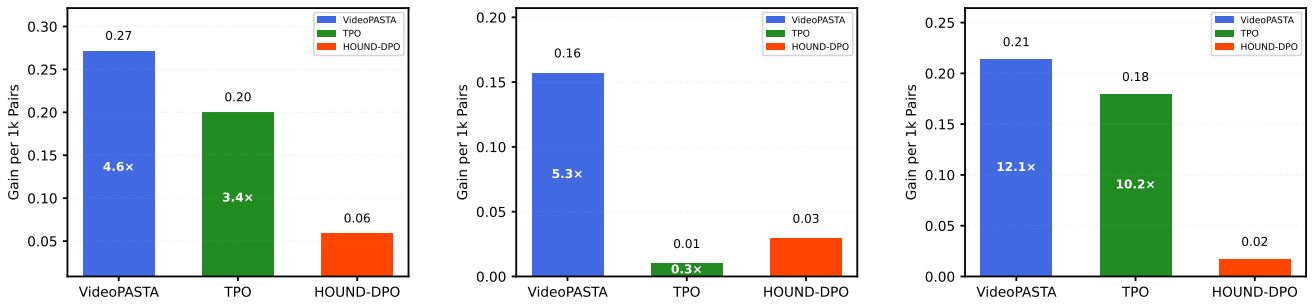
On VideoMME, VideoPASTA achieves $\mathcal{G}_{\text{VideoPASTA}} = 0.27$ points of improvement per 1k pairs, compared to TPO's $\mathcal{G}_{\text{TPO}} = 0.20$ and Hound-DPO's $\mathcal{G}_{\text{Hound-DPO}} = 0.06$, representing $1.4\times$ and $4.5\times$ higher efficiency, respectively. Even higher differences emerge on MVBench ($\mathcal{G}_{\text{VideoPASTA}} = 0.16$ points per 1k pairs, $16\times$ more efficient than TPO's $\mathcal{G}_{\text{TPO}} = 0.01$ and $5.3\times$ more efficient than Hound-DPO's $\mathcal{G}_{\text{Hound-DPO}} = 0.03$) and NeXTQA ($\mathcal{G}_{\text{VideoPASTA}} = 0.21$ points per 1k pairs, $10.2\times$ more efficient than TPO's $\mathcal{G}_{\text{TPO}} = 0.18$ and $12.1\times$ more than Hound-DPO's $\mathcal{G}_{\text{Hound-DPO}} = 0.02$). These substantial efficiency gaps demonstrate that structured adversarial examples targeting specific failure modes provide significantly more informative learning signals than generic preference data.



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

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

Figure 6. Performance vs. Number of Preference Pairs across six benchmarks. **VideoPASTA achieves superior results with only 7k pairs** compared to TPO [28] (10k pairs) and and Hound-DPO [70] (17k pairs). Note the **consistent upward trajectory** for VideoPASTA vs. the performance degradation of Hound-DPO on MLVU and LongVideoBench at higher pair counts, highlighting the importance of **targeted adversarial examples** over mere data quantity.



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

Figure 7. **Information gain analysis across three representative benchmarks**. Each bar represents performance improvement per 1K preference pairs, calculated as $(final\_score - baseline\_score)/num\_pairs\_in\_thousands$. VideoPASTA demonstrates dramatically higher learning efficiency compared to competing methods, up to 16× more efficient than TPO on MVBench and 12.1× more efficient than Hound-DPO on NeXTQA highlighting the impact of targeted adversarial examples in creating informationally dense learning signals.

Stability of Learning Trajectories. Interestingly, we observe divergent trends for LLaVA-Hound-DPO across several benchmarks, where performance actually degrades at higher example counts (particularly evident in MLVU and LongVideoBench). This suggests that preference pairs generated using proprietary models without explicit targeting of failure modes may introduce noise that compromises alignment as training progresses. In contrast, VideoPASTA's targeted approach maintains consistent improvement trajectories, confirming that structuring adversarial examples around specific weaknesses yields more robust optimization.

Quality Over Quantity. While our previous ablation study (Figure 5) demonstrates that VideoPASTA benefits from additional data within its structured framework, these results highlight the critical distinction between data quantity and quality. Even when all methods benefit from more examples, the rate of improvement per example for VideoPASTA significantly outpaces alternatives, allowing it to achieve superior performance with 30% fewer examples than TPO and 59% fewer than LLaVA-Hound-DPO. This efficiency shows that

systematically targeting failure modes through structured adversarial examples creates a more effective learning signal than generic preference data, regardless of quantity.

A.3. Dataset Overview

A.3.1. Dataset Statistics

Starting with 3000 videos from ActivityNet [64], 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 $r_{\text{crossframe}}$, representing spatial, temporal, and cross-frame failure modes, respectively. Theoretically, this setup yields:

$$\begin{aligned} N_{\text{potential}} &= |V| \times |Q| \times |R^-| \\ &= 3000 \times 10 \times 3 = 90,000 \end{aligned} \tag{5}$$

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 [61] verification using the prompt template given in Figure 13. Each preference pair must satisfy three criteria:

1. The preferred 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:

$$N_{\text{final}} = N_{\text{potential}} \times r_{\text{retention}} \approx 90\,000 \times 0.078 \approx 7,020, \tag{6}$$

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.

A.3.2. Dataset Samples

Figure 8 illustrates key examples from our preference dataset that demonstrate how VideoPASTA targets specific failure modes in video-language understanding. These examples were carefully curated to challenge different aspects of video comprehension while maintaining clear distinctions between preferred and dispreferred responses.

Spatial Misalignment. The boat counting example demonstrates our approach to spatial reasoning. While the dispreferred 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-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 dispreferred 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 dispreferred response mistakenly interprets superficial visual changes, such as a close-up of a face, as significant narrative shifts, whereas the preferred 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 [61] to ensure that dispreferred 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.

A.4. Qualitative Examples

We present several representative examples that demonstrate how VideoPASTA improves video understanding across various scenarios. Figure 9 illustrates three key aspects of our model's capabilities in handling complex video content.

First, in the camera advertisement sequence, while Qwen2.5-VL [3] 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

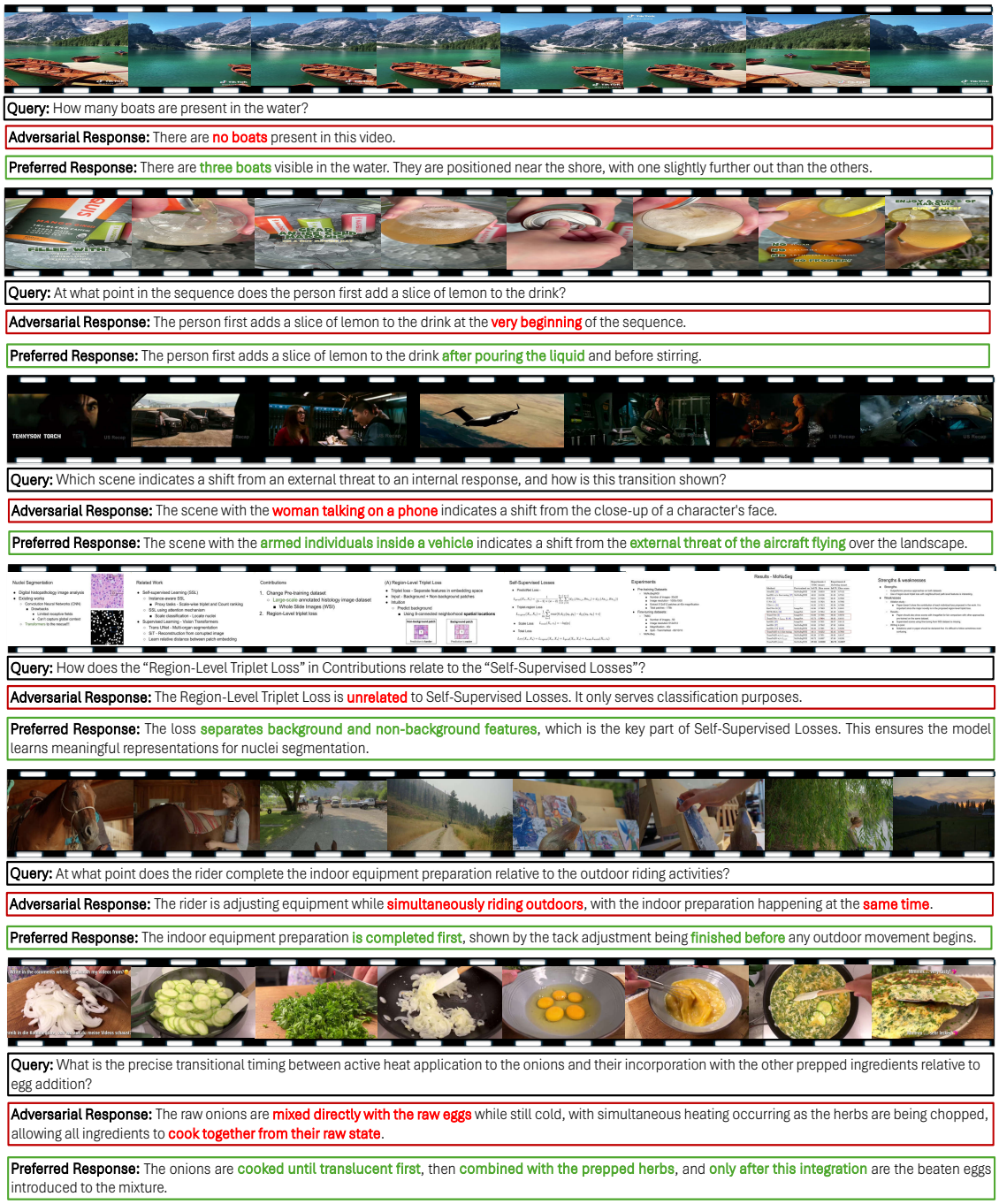


Figure 8. **Examples from our preference dataset demonstrating diverse failure modes targeted by VideoPASTA**. Each row shows a video sequence with its corresponding query and preferred and adversarial responses. The adversarial samples cover *spatial misalignment* (counting objects in scenes), *temporal 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 preferred responses maintain accurate spatial-temporal alignment with the video content.

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

Query: What is the narrative progression of this video, and what key transformation occurs between the initial indoor camera moments and the final outdoor scenes?

Qwen2.5VL: This is a **travel documentary** showing **unrelated clips** of someone looking at a camera, **randomly followed** by tourist spots and tunnel visits.

VideoPASTA: This is a **Nikon camera advertisement** that shows the journey from **photographer to photograph**, beginning with camera handling moments that **result into** the resulting landscape and adventure photography.

Query: What is the progression of emotional states and actions shown in the muscular character's response to Bugs Bunny's basketball challenge?

Qwen2.5VL: The character becomes enraged **immediately after** seeing Bugs Bunny, then starts grunting while exercising with weights, followed by his jumping attempt, with the steam effects appearing in the video **as a result of** his physical workout.

VideoPASTA: The character **first shows irritation** during the basketball game, which **escalates into** visible anger with grunting, then **transitions to** explosive rage, and **finally concludes** in an airborne leap of fury after Bugs Bunny's taunt.

Query: In what order are the green chilies prepared and coated before frying?

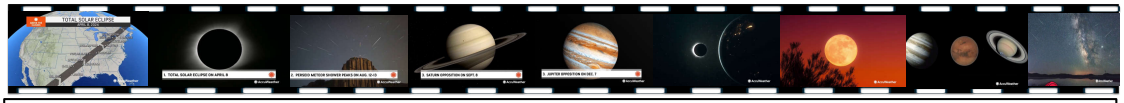
Qwen2.5VL: The chilies are **first dipped in egg wash**, then **slit and cleaned**, followed by **coating in breadcrumbs**, and then **dipped in flour** before being fried until golden brown.

VideoPASTA: The chilies are **first cleaned and slit**, then **coated in the flour mixture**, followed by **dipping in the egg wash**, and **finally covered in breadcrumbs** before being deep fried until golden brown.

Query: What is the sequence of the athlete's emotional and physical reactions from the moment he completes his jump until the final result?

Qwen2.5VL: After landing, he **immediately jumps up celebrating**, then **waits for the measurement**, followed by **lying down in the sand** to rest, before finally showing disappointment at his result.

VideoPASTA: The athlete **first lies motionless** in the sand pit, then **slowly sits up**, followed by a knowing **smile** at his performance, which is **confirmed by his celebration** when seeing the 8.58m measurement displayed.



Query: What visual changes are shown in the sky during the total solar eclipse footage?

Qwen2.5VL: The sun is **suddenly blocked** by the moon, showing only a bright ring around the edges, while the sky remains **partially lit** throughout the event.

VideoPASTA: The bright sun is **gradually covered** by the moon, creating a **dark circular shadow** with a glowing white ring around it, and stars become visible in the darkened sky.

Query: How does the girl assist the boy in levitating the feather, and what evidence supports this assistance?

Qwen2.5VL: The girl assists the boy by simply **watching him** attempt the spell. There is **no clear evidence** of her providing specific guidance or help with the wand movement.

VideoPASTA: The girl assists the boy by **demonstrating the correct wand movement**. Evidence includes her **focused expression** and the **feather's subsequent levitation after her guidance**.

Figure 9. **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.

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.

A.5. 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 10–13) 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 10. 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 11. 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 12. 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 **preferred** 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 *preferred* 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 preferred.

For each query and its four responses:

1. **Sanity-check the preferred 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 preferred response or fails to demonstrate a clear misalignment, discard it.

Figure 13. Prompt template for validating one preferred 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 14. 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 15. 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]
Preferred 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 preferred 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 16. Prompt used for evaluating whether adversarial examples correctly target their claimed failure modes.