TEMPORAL PREFERENCE OPTIMIZATION OF LARGE MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent advancements in video large multimodal models (video-LMMs), accurate temporal grounding remains a key challenge. In this work, we introduce **Temporal Preference Optimization** (**TPO**)—a post-training framework that unlocks superior temporal reasoning in video-LMMs without requiring human annotations. TPO enables preference modeling by manipulating video inputs to generate contrastive responses, ensuring that preferred responses are more temporally grounded than dis-preferred ones. Through preference learning, TPO enhances the model's capability for more comprehensive video understanding with better temporal reasoning. Extensive experiments on LongVideoBench, MLVU, and Video-MME demonstrate that TPO significantly improves temporal grounding across multiple video-LMMs. Notably, LLaVA-Video-TPO achieves state-of-the-art performance among 7B models on Video-MME, establishing TPO as a scalable and effective solution for advancing temporal understanding in video analysis.

1 INTRODUCTION

Recent advances in video large multimodal models (video-LMMs) (Wang et al., 2024b; Achiam et al., 2023; Reid et al., 2024) represents a significant step toward general video understanding. While image-based LMMs (Hong et al., 2024; Bai et al., 2023; Lu et al., 2024) primarily focus on spatial reasoning, video-LMMs face the additional complexity of modeling temporal dependencies—a critical aspect for capturing the dynamic nature of video content.

Most existing video-LMMs are trained through supervised finetuning with video–question–answer pairs, without explicit mechanisms for temporal grounding. Consequently, temporal alignment is only acquired implicitly, and models often struggle to localize the precise moments that support their responses (Chen et al., 2024a; Zhang et al., 2024d). Recent efforts (Ren et al., 2024; Chen et al., 2024b; Li et al., 2023a; Huang et al., 2024; Wang et al., 2024a) have sought to improve grounding by enriching textual responses with structured temporal information and incorporating explicit segment-level annotations into training. While providing stronger supervision, it relies on additional temporal annotations, which are costly to obtain and difficult to scale to large datasets.

We introduce Temporal Preference Optimization (TPO), a post-training framework that enhances temporal grounding in video-LMMs without requiring manual annotations. TPO generates contrastive supervision by prompting a model with the same query on both original and corrupted videos: responses from relevant frames are treated as preferred, while those from irrelevant or incomplete frames are dispreferred, forming a natural preference hierarchy. A lightweight post-filtering step removes noisy or ambiguous samples, yielding a clean preference dataset. This dataset is then used to refine temporal grounding through Direct Preference Optimization (Rafailov et al., 2024), which improves temporal reasoning while preserving pretrained knowledge. By automatically injecting temporal preferences through simple input transformations, TPO provides a scalable and robust solution for advancing temporal reasoning tasks.

We conducted extensive experiments on three challenging multimodal video understanding benchmarks, and the results clearly demonstrate that TPO significantly enhances the temporal grounding capabilities of video-LMMs. Specifically, TPO achieves performance gains of 2.9% on LongVideoBench (Wu et al., 2024), 3.1% on MLVU (Zhou et al., 2024a), and 2.5% on Video-MME (Fu et al., 2024), when applied to the strong base model LongVA-7B (Zhang et al., 2024b).

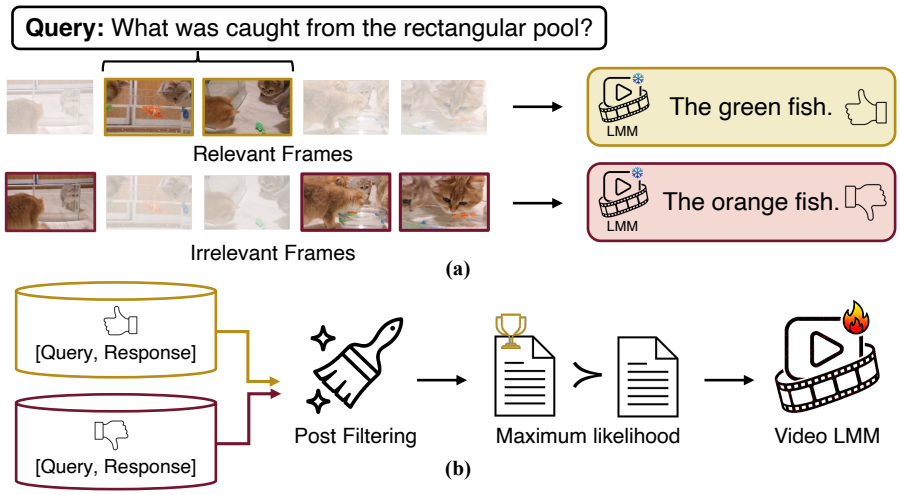


Figure 1: **Temporal Preference Optimization.** (a) The model self-generates preference data by producing contrastive responses to well-grounded versus perturbed (irrelevant or incomplete) video clips. (b) An LLM-based post-filtering step removes noisy or misaligned samples, and the refined preference data is then used in preference optimization. Through this self-improvement process, the model learns to favor temporally consistent responses, ultimately strengthening temporal reasoning.

Furthermore, even when integrated with the state-of-the-art large-scale pretrained video-LMM, LLaVA-Video, TPO still delivers a 2.3% improvement, establishing LLaVA-Video-TPO as the top-performing 7B model on the Video-MME benchmark.

2 Temporal Preference Optimization

While prior works focus primarily on aligning LLM outputs with human preferences, our approach uniquely aligns model outputs with intrinsic temporal preferences in videos. To achieve this, we propose Temporal Preference Optimization (TPO) (Fig. 2), a framework that enhances video-LMMs' temporal reasoning by explicitly incorporating temporal modeling into the optimization process. TPO generates preference pairs through contrasts between meticulously manipulated video inputs (Sec. 2.1). To further enhance the preference data quality, we introduce a rule-based post-filtering step (Sec. 2.2). Finally, Direct Preference Optimization (Sec. 2.3) is leveraged to optimize the model towards temporally preferred outputs without compromising its original pretrained capabilities.

2.1 Temporal Preference Modeling

Query Generation. Given a video $\mathbf{V}$, we first sample a segment containing a set of frames S_a, which may be a subset of the video or the entire sequence of frames. To generate descriptive context, we employ CogVLM2 (Hong et al., 2024), an image-based LMM, to generate captions for each frame in S_a. These captions serve as the foundation for constructing targeted questions. To ensure diversity and relevance, we design multiple question types and use a structured question-generation prompt to incorporate the generated captions, as shown in Fig. 9 (Appendix). This prompt is then processed by a LLM (GPT-4o-mini) to produce a set of candidate questions specifically tailored to the sampled video frames, resulting in a set of questions S_q. This approach ensures that the generated questions are contextually relevant that allows precise control over subsequent response generation.

Preferred Response Generation. Preferred responses in the curated dataset are expected to be strongly grounded in the corresponding temporal content. To achieve this, we use the question set S_q along with their corresponding frame set S_a as input to the video-LMM. Given the provided video frames are highly relevant to the query, we create conditions that maximize the likelihood of generating a high-quality, temporally grounded response. This process guarantees that the preferred responses align with the ideal characteristics for effective temporal grounding in video-LMMs.

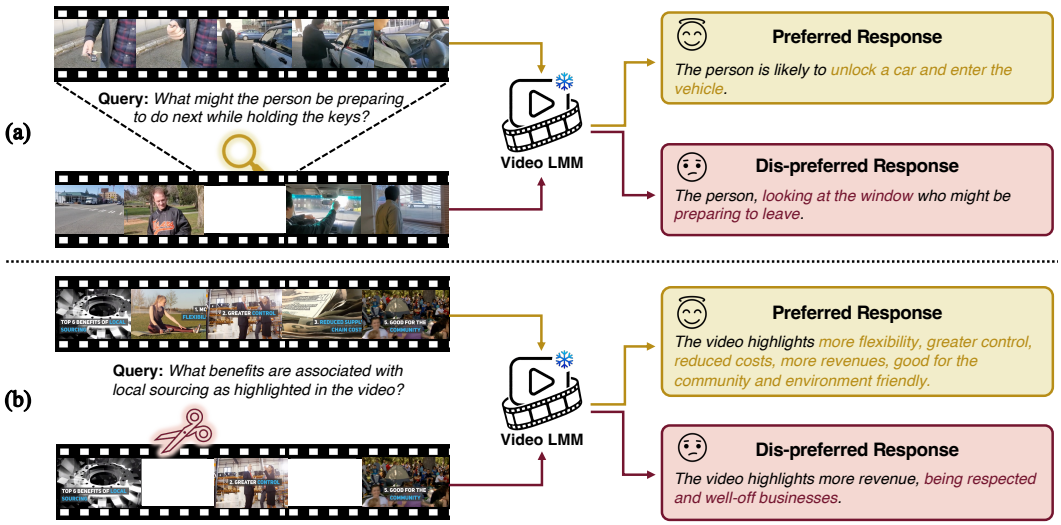


Figure 2: To ensure strong temporal grounding, we generate preferred responses using the full relevant frame set S_a. For dis-preferred responses, we introduce: (a) **Generation with Irrelevant Information**, where all relevant frames are excluded. (b) **Generation with Incomplete Information**, where only a partial subset of relevant frames is used. These manipulated clips create contrastive response pairs, highlighting differences between well-grounded and manipulated video content. This contrast serves as a learning signal to enhance the model's temporal reasoning.

Dis-Preferred Response Generation. In our preference dataset, dispreferred responses are those that the model is trained to avoid—outputs that fail to temporally localize the evidence in the video. These serve as hard negative examples for temporal reasoning, highlighting cases where the model struggles to align its predictions with the actual video content. To generate these dis-preferred responses, we manipulate the video inputs to simulate imperfect temporal grounding. As illustrated in Fig. 2, we introduce two strategies for constructing the input frame set S_b used in dis-preferred response generation:

(a) Generation with Irrelevant Information: To simulate an extreme failure case where the model misses all relevant frames, we construct S_b by excluding the relevant frame set S_a and instead sampling from the remaining frames of the video. This ensures that S_b contains only irrelevant content, forcing the model to generate a response based on unrelated visual information.

(b) Generation with Incomplete Information: Simulating the model can only consume partial relevant information, S_b is randomly sampled as a subset of S_a. This setup introduces gaps in the temporal context, making it harder for the model to fully comprehend the key event in the query.

Unlike preferred responses, which are grounded in fully relevant video segments, these manipulated setups introduce ambiguity and noise by partially or completely omitting critical visual content. As a result, the model is forced to rely on incomplete or misleading information, making temporal reasoning errors and hallucinations more likely. This intentional contrast between preferred and dis-preferred responses serves as a strong learning signal, helping refine the model's ability to distinguish and accurately localize events in time, ultimately enhancing its temporal reasoning capabilities.

2.2 LLM-BASED POST-FILTERING

Although we design the preferred responses to be higher quality than the dis-preferred responses, this distinction is not always guaranteed due to the limitations of the base video-LMMs. In some cases, errors in response generation may lead to misaligned preference pairs, where the preferred response contains noise or the dis-preferred response is of better quality than expected.

To enhance data quality and reduce noise, we introduce a post-filtering pipeline with an LLM (GPT-4o-mini). Specifically, we provide the model with the key frame captions of S_a, along with their corresponding queries and preference data pairs, and instruct it to filter out samples that meet

predefined criteria (detailed prompts are shown in Fig. 10 in the Appendix). The filtering rules target cases where: 1) The dis-preferred response is of higher quality than the preferred response. 2) The preferred response is factually incorrect or misaligned with the video content. 3) The query is ambiguous, making preference ranking unreliable. By incorporating this post-filtering step, we effectively eliminate potential noisy cases, resulting in a refined, higher-quality dataset that better supports effective model optimization and improves temporal grounding performance.

2.3 Training Objective

The generated preference dataset is leveraged to optimize the temporal grounding capabilities of video-LMMs using Direct Preference Optimization (DPO) (Rafailov et al., 2024), selected for its robustness and effectiveness in preference-based learning.

Given the preference dataset D (V, q, r^+, r^-) and a video-LMM π_θ, the DPO loss is defined as:

$$L_{DPO}(\pi_\theta; \pi_{ref}) = -E_{(V,q,r^+,r^-)\sim\mathcal{D}} \left[log\sigma(\beta(\log \frac{\pi_\theta(r^+|V,q)}{\pi_{ref}(r^+|V,q)} - \log \frac{\pi_\theta(r^-|V,q)}{\pi_{ref}(r^-|V,q)})) \right] \quad (1)$$

where σ is the sigmoid function. This objective drives the model to assign higher probabilities to preferred outputs, aligning its behavior more closely with human judgments, while preventing the model from deviating too much from its pretrained distribution.

To better align the model with the preferred responses, we incorporate a supervised fine-tuning objective into the DPO training framework. This combined objective is controlled by the hyperparameter α, following (Chen et al., 2021; Deng et al., 2024; Chen et al., 2023).

$$L_{SFT}(\pi_\theta) = -E_{(V,q,r^+,r^-)\sim\mathcal{D}} \log \pi_\theta(r^+|V,q) \quad (2)$$

$$L(\pi_\theta; \pi_{ref}) = L_{DPO} + \alpha L_{SFT} \quad (3)$$

3 Experiments

3.1 Experimental Settings

Evaluation Benchmarks We evaluate TPO and baselines on three widely recognized benchmarks in multimodal video understanding. **Video-MME** (Fu et al., 2024) offers a comprehensive multi-modal evaluation across diverse video lengths, spanning from 11 seconds to 1 hour. **LongVideoBench** (Wu et al., 2024) emphasizes reasoning tasks within extended video contexts. **MLVU** (Zhou et al., 2024a) supports multitask evaluation specifically designed for long-form video understanding.

Models We test the effectiveness of TPO on two popular video-LMMs, LongVA-7B (Zhang et al., 2024b) and LLaVA-Video-7B (Zhang et al., 2024e), deriving the following models:

- LongVA-TPO: optimized based on LongVA-7B (Zhang et al., 2024b), a capable video-LMM with the long-context video understanding capability transferred from language.
- LLaVA-Video-TPO: optimized based on LLaVA-Video-7B (Zhang et al., 2024e), the current state-of-the-art 7B video-LMM.

In the main text, all ablation studies and analyses are conducted using LongVA-TPO by default; additional ablation results are provided in the appendix.

Implementation Details For the video source in preference dataset generation, we manually curated 200 keywords to retrieve 8k videos from the internet, ensuring diversity and coverage. For each video, we sampled 32 frames. To simulate irrelevant information, we randomly selected 4 consecutive frames as S_a and used the remaining 28 frames as S_b. To simulate incomplete information, we instead used all 32 frames as S_a and uniformly removed 16 frames to form S_b. Using this pipeline, we created 10k preference data pairs for LongVA-TPO using our established pipeline. For LLaVA-Video-TPO, we employ a subset of the original LLaVA-Video-178K dataset, which was used for supervised fine-tuning (SFT), to generate TPO data, resulting in a total of 10K preference data pairs.

Model	Size	LongVideo Bench	MLVU (M-avg)	Video-MME			
				Short	Medium	Long	Average
GPT-4o	-	66.7	64.6	80.0/82.8	70.3/76.6	65.3/72.1	71.9/77.2
Video-LLaVA	7B	39.1	47.3	45.3/46.1	38.0/40.7	36.2/38.1	39.9/41.6
PLLaVA	7B	40.2	-	-	-	-	-
Qwen-VL-Max	-	-	42.2	55.8/57.6	49.2/48.9	48.9/47.0	51.3/51.2
ShareGPT4Video	8B	39.7	46.4	48.3/53.6	36.3/39.3	35.0/37.9	39.9/43.6
InternVL-Chat-V1.5	20B	51.2	50.4	50.7/52.4	60.2/61.7	46.4/49.1	45.6/46.6
VideoChat2	7B	39.3	47.9	48.3/52.8	37.0/39.4	33.2/39.2	39.5/43.8
LongLLaVA	7B	-	56.3	61.9/66.2	51.4/54.7	45.4/50.3	52.9/57.1
Video-CCAM	14B	-	63.1	62.2/66.0	50.6/56.3	46.7/49.9	53.2/57.4
NVILA	7B	57.7	70.1	75.7/77.6	62.2/69.0	54.8/63.3	64.2/70.0
Qwen2-VL	7B	55.6	-	-	-	-	63.3/69.0
Apollo	7B	58.5	70.9	-	-	-	61.3/63.3
MiniCPM-o-2.6	7B	-	-	75.4/78.3	63.9/69.1	52.2/56.3	63.9/67.9
LongVILA	7B	57.1	-	69.0/72.9	58.3/64.9	53.0/57.4	60.1/65.1
LiveCC	7B	-	-	74.8/76.6	63.9/70.3	53.7/64.1	64.1/70.3
Qwen2.5-VL	7B	56.0	70.2	-	-	-	65.1/**71.6**
LongVA-7B	7B	51.3	58.8	61.1/61.6	50.4/53.6	46.2/47.6	52.6/54.3
LLaVA-Video-7B	7B	58.2	70.8	-	-	-	63.3/69.7
LongVA-TPO	7B	54.2	61.7	63.1/66.6	54.8/55.3	47.4/47.9	55.1/56.6
LLaVA-Video-TPO	7B	**60.1**	**71.1**	**76.8/78.7**	**64.6/69.4**	**55.4/66.4**	**65.6**/71.5

Table 1: Results on LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2024a) and Video-MME (Fu et al., 2024) compared with state-of-the-art models. The Video-MME results are presented in the format "w/o subs / w/ subs".

The model is trained on 8 Nvidia A100 80GB GPUs, with a batch size of 64. For the preference optimization on LongVA, we set the KL-divergence weight $\beta = 0.3$ and the SFT loss weight $\alpha = 0.5$, while for LLaVA-Video, we set the KL-divergence weight $\beta = 0.2$ and the SFT loss weight $\alpha = 1$. To ensure information consistency, we use the same sampled 32 frames for both data generation and model training. We train the model on our curated data for 1 epoch. It takes about 4 hours for TPO to perform on LongVA-7B with a learning rate of $4e^{-6}$ and also about 4 hours for LLaVA-Video-7B with a learning rate of $3e^{-7}$. During data preparation, we employ the GPT-4o-mini (text-only input) for question curation and post-filtering. This choice balances cost-effectiveness with efficiency, facilitating a streamlined and scalable data processing workflow.

3.2 Results

The comparisons between TPO and current state-of-the-art video-LMMs on LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2024a), and Video-MME (Fu et al., 2024) are presented in Table 1. With the introduction of TPO, both the LongVA-TPO and LLaVA-Video-TPO models significantly outperform their corresponding baselines 2.5% and 2.3% on the Video-MME benchmark, demonstrating the efficacy of our TPO pipeline. After TPO on LLaVA-Video-7B, our LLaVA-Video-TPO model outperforms all 18 baseline models in the table, including the concurrent work, as well as several 14B and 20B models, achieving state-of-the-art results on video understanding. The original LongVA model performed worse than Video-CCAM (Fei et al., 2024) and LongLLaVA (Yin Song and Chen Wu and Eden Duthie, 2024) on the Video-MME benchmark. However, after incorporating TPO, it successfully outperformed these competitive baselines on Video-MME. Overall, LLaVA-Video-TPO achieves the strongest 7B model on Video-MME, setting a new state-of-the-art performance on video comprehension.

Besides, we compare TPO against three different training strategies on LongVA in Table 2:

- SFT_{Self}: Supervised fine-tuning using the self-generated data. For a fair comparison, we utilize the same preferred response in our curated preference dataset to optimize LongVA.
- SFT_{LLM}: Supervised fine-tuning using the LLM-generated data. Following the commonly used data curation pipeline (Chen et al., 2024a; Zhang et al., 2024d). We employ LLM (GPT-4o-mini) to

Model	LongVideoBench	MLVU (M-avg)	Video-MME			
			Short	Medium	Long	Average
LongVA-7B	51.3	58.8	61.1/61.6	50.4/53.6	46.2/47.6	52.6/54.3
+ SFT_{Self}	52.7	58.9	62.6/**67.7**	52.4/52.7	46.8/47.4	53.9/55.9
+ SFT_{LLM}	53.1	59.9	**63.7**/64.9	52.6/54.3	46.3/47.9	54.2/55.7
+ Hound-DPO†	52.8	59.1	62.2/65.8	52.4/54.8	46.1/46.3	53.6/55.6
+ Hound-DPO*	52.6	59.3	63.1/65.9	50.8/54.7	47.2/47.0	53.7/55.9
LongVA-TPO	**54.2**	**61.7**	63.1/66.6	**54.8/55.3**	**47.4/47.9**	**55.1/56.6**

Table 2: Results of LongVA-TPO on LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2024a) and Video-MME (Fu et al., 2024) benchmarks compared to baseline methods mentioned in 3.2. The Video-MME results are presented in the format "w/o subs / w/ subs". The results for LongVA and LongVA+Hound-DPO† (Zhang et al., 2024c;b) are based on publicly available checkpoints, while LongVA+Hound-DPO* is reproduced using our collected video datasets.

generate responses given the query and the video captions, which are subsequently used to perform supervised fine-tuning on LongVA. We use the same video data as TPO for fair comparison.

- Hound-DPO (Zhang et al., 2024c): Applying Direct Preference Optimization (DPO) (Rafailov et al., 2024) to video-LMMs, Hound-DPO leverages LLM to rate preference data, resulting in a dataset of 17k samples. In contrast, TPO relies on a smaller, self-generated preference dataset, offering a more streamlined alternative. Besides, to ablate the data source's effect, we also implement Hound-DPO based on our collected dataset with the same data scale.

The results consistently indicate that LongVA-TPO achieves superior performance, with improvements of 2.9%, 3.1%, and 2.5% on LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2024a), and Video-MME, respectively. These findings underscore TPO's capacity to enhance the general video understanding capabilities of a pre-trained video-LMM.

Compared to SFT_{Self}, LongVA-TPO achieves a consistent performance gain of 1.2% to 2.8% by utilizing carefully designed temporal preference pairs. Furthermore, LongVA-TPO outperforms SFT_{LLM}, demonstrating the effectiveness and stability of our self-training paradigm. When compared to Hound-DPO (Zhang et al., 2024c), LongVA-TPO achieves a significant performance improvement by modeling temporal preference priors. However, LongVA-TPO achieves comparable performance compared to SFT methods on the Video-MME-short subset, which aligns with TPO's design focus—enhancing temporal reasoning in longer videos.

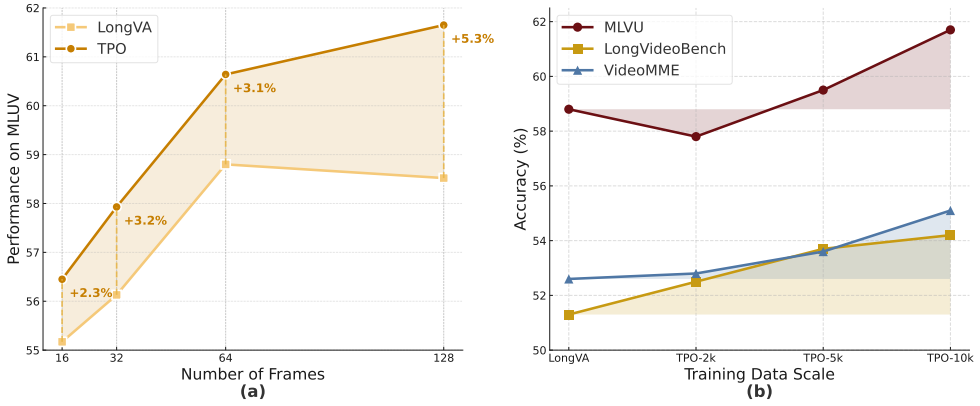


Figure 3: (a) Performance of LongVA-TPO vs. LongVA on MLVU across input lengths. LongVA-TPO consistently benefits from increased input length, whereas LongVA's performance declines when inputs exceed 64 frames. (b) The consistently improved performance of LongVA-TPO when scaling the data. The performance on the Video-MME benchmark is evaluated without subtitles.

3.3 Analysis

Effect of Input Frame Count We evaluate the performance of LongVA-TPO and LongVA across input lengths from 16 to 128 frames (Fig. 3(a)). While LongVA's performance drops at 128 frames compared to 64, LongVA-TPO continues to improve with longer sequences, consistently benefit from longer input. This demonstrates LongVA-TPO's robustness to extended inputs and its capacity to localize relevant information within long sequences, underscoring the effectiveness of TPO.

Effect of Dataset Sizes Scalability is a critical metric in the evaluation of algorithms in the era of large-scale models, reflecting an algorithm's performance as data volume expands. To assess this, we evaluate LongVA-TPO on across incremental sizes of 2k, 5k, and 10k. As shown in Fig. 3(b), performance improves consistently with larger datasets across all three benchmarks, demonstrating superior scalability. This pattern highlights TPO's robustness and adaptability in larger data contexts, suggesting its potential to deliver enhanced results when scaled to larger datasets.

Effect of Post-Filtering As a critical component of the TPO framework, post-filtering effectively reduces noise and enhances data quality. After post-filtering, we manually reviewed 200 preference pairs and found that 96.5% satisfied our criteria—where the preferred response was both more appropriate than the dispreferred one and accurately answered the question based on the video content. To further assess its impact, we conducted experiments comparing the performance of LongVA-TPO with and without post-filtering. The results, presented in Fig. 4 (a), demonstrate that post-filtering consistently improves performance across multiple benchmarks.

Effect of Different Data Mix Ratio In TPO, we design two different kinds of manipulated schema for the dis-preferred response generation. To evaluate their individual effects and the benefit of combining them, we conducted an ablation study while keeping the dataset size fixed. We tested mixing ratios of 10:0, 8:2, 5:5, 2:8, and 0:10 between generation with irrelevant and incomplete information. As shown in Fig. 4(b), performance peaks with a 5:5 ratio. The balanced data distribution effectively integrates different type of temporal information, leading to superior performance.

Needle-in-a-Haystack (NIAH) The NIAH task challenges models to detect rare events within long videos. Following Zhang et al. (2024b), we frame the task using the same five image-based question answerings (QAs), where images are embedded within a 3-hour-long video, and the model is tasked with answering the corresponding image QA. We report results in Fig. 5. While LongVA, optimized for long-context processing, significantly outperforms LLaVA-NeXT-Video (Zhang et al., 2024d) on

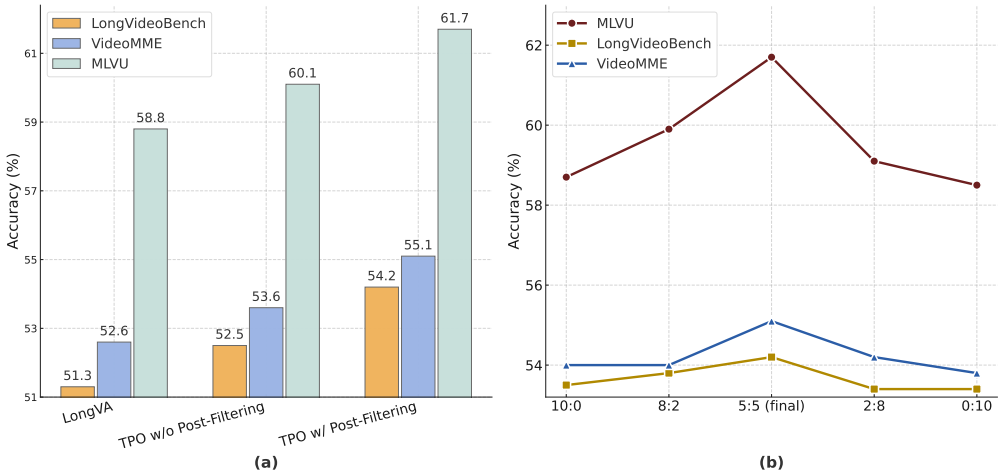


Figure 4: (a) Post-filtering consistently improves performance across multiple benchmarks on LongVA. (b) Performance of TPO across different training data mix ratios, varying the proportion of negative responses generated from incomplete versus irrelevant video segments.

Figure 5: Performance comparison of LongVA and LongVA-TPO on the needle-in-a-haystack task across varying input video lengths (horizontal axis) and temporal depths (vertical axis). Heatmaps indicate improved temporal grounding capability of LongVA-TPO.

the NIAH task (refer to Fig. 4 in Zhang et al. (2024b)), our LongVA-TPO model still demonstrates superior performance, achieving even better results in long-context temporal localization.

3.4 Qualitative Analysis

The qualitative analysis of LongVA-TPO and LongVA on two videos from the Video-MME benchmark is provided in Fig. 6. In the first example, which involves a temporal localization and OCR task, LongVA-TPO accurately localizes the relevant information within the video and providing the correct answer to the OCR question. In the second example, a video discussing the Moon's formation, LongVA misinterprets the video content by relating it to the Earth's formation. In contrast, LongVA-TPO successfully comprehends and captures the key details of the video's content.

4 Related Work

Video Large Multimodal Models (video-LMMs) Recently, significant efforts have extended large language models (Achiam et al., 2023; Reid et al., 2024) into the visual domain, leading to the development of both proprietary (Achiam et al., 2023; Reid et al., 2024) and and open-source video-LMMs (Wang et al., 2024b; Liu et al., 2024a; Li et al., 2024b; Shen et al., 2024; Lin et al., 2023; Chen et al., 2024c; Yao et al., 2024; Fu et al., 2025; Abdin et al., 2024; Laurençon et al., 2024; Chen et al., 2025c;a; Bai et al., 2025). Early work emphasized video-text instruction-tuning datasets (Chen et al., 2024a; Zhang et al., 2024e; Liu et al., 2023; Park et al., 2023; Zhang et al., 2024d), but their reliance on synthetic captions limits their effectiveness in capturing visual-temporal dynamics. Other studies have focused on extending pretrained video-LMMs for long contexts (Zhang et al., 2024b; Liu et al., 2024d; Yin Song and Chen Wu and Eden Duthie, 2024; Liu et al., 2024b;c; Shu et al., 2024; Islam et al., 2025), while multimodal interleaved datasets (Li et al., 2024c; Lin et al., 2024) and mixed training strategies (Zohar et al., 2024b; Li et al., 2024a) have been explored to enhance video-LMM performance. However, the post-training stage for video-LMMs remains underexplored. Recent efforts like LLaVA-Hound (Zhang et al., 2024c) builds preference datasets by ranking model outputs with LLM but fall short in leveraging the temporal information inherent in video. In contrast, our work pioneers post-training strategies that explicitly incorporate temporal priors.

Temporal grounding is crucial for comprehending the video modality, particularly in long-form videos. Prior work has explored diverse strategies, including dense captioning (Wang et al., 2021; Yeung et al., 2018; Yang et al., 2023), highlight detection (Lei et al., 2021; Moon et al., 2023), and temporal video grounding (Gao et al., 2017; Yuan et al., 2019; Xiao et al., 2024), among others. Recent advancements have introduced temporal-aware designs in video-LMMs (Ren et al., 2024; Chen et al., 2024b; Li et al., 2023a; Huang et al., 2024; Wang et al., 2024a) and agentic systems with temporal grounding capabilities have emerged (Wang et al., 2025a). Besides, reinforcement learning (Shao et al., 2024) has been recently applied to temporal grounding Chen et al. (2025b); Wang et al. (2025b;c); Li et al. (2025). Despite providing explicit temporal supervision, these methods require additional temporal annotations, which are costly and often impractical at large scales.

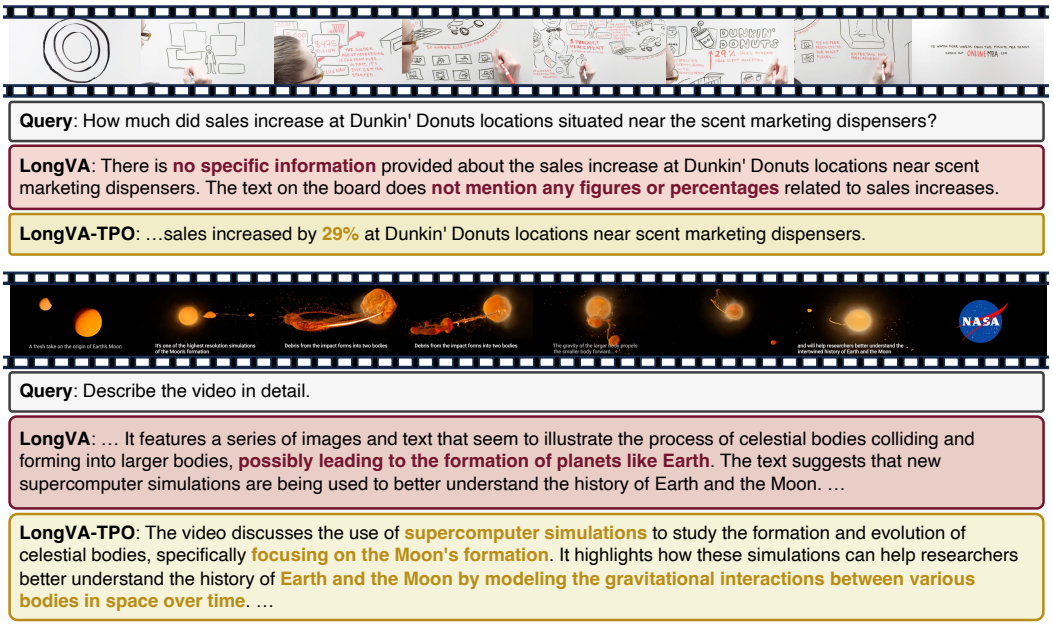


Figure 6: Qualitative comparison between LongVA-TPO and LongVA on videos from Video-MME.

Proximal Policy Optimization (Ouyang et al., 2022; Stiennon et al., 2020; Ziegler et al., 2019) and Direct Preference Optimization (Rafailov et al., 2024) are two widely used implementations of Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Ziegler et al., 2019), serving as key algorithms in preference learning and post-training. In image-LMMs, Sun et al. (2023) improved model's visual capability by incorporating image captions into the reward model. Similarly, Ahn et al. (2024b) fine-tuned multimodal models using Reinforcement Learning from AI Feedback. Other approaches distilled GPT-4V's preferences directly from sampled responses (Li et al., 2023b; Gunjal et al., 2024) or leveraged text as an intermediate modality, using captions and other descriptive information to extract preferences for images (Zhao et al., 2023) and videos (Zhang et al., 2024c). Furthermore, Pi et al. (2024); Zhou et al. (2024b); Deng et al. (2024) advanced preference learning in image-LMMs by curating preference data through image input manipulation.

Self-Training in Foundation Models To reduce reliance on large annotated datasets, several works have explored self-improvement and self-training methods (Huang et al., 2022; Ho et al., 2022). Zelikman et al. (2022) introduced Self-Taught Reasoners, which leverage generated chain-of-thought rationales to enhance LLMs' complex reasoning capabilities. For images, BPO (Pi et al., 2024), STIC (Deng et al., 2024) and POVID (Zhou et al., 2024b) improve image-LMMs responses by incorporating visual priors. For videos, Video-STaR (Zohar et al., 2024a) uses existing labels as weak supervision while Ahn et al. (2024a) explores iterative self-improvement in preference optimization.

5 Conclusion

We introduced Temporal Preference Optimization (TPO), a scalable post-training framework that enhances temporal grounding in video-LMMs. By contrasting between the preference responses from the well-grounded and manipulated video clips, TPO effectively captures the intricate temporal dependencies required for video understanding. Extensive experiments across three challenging benchmarks—LongVideoBench, MLVU, and Video-MME—demonstrated TPO's robust improvements, achieving state-of-the-art performance. By integrating multi-granularity temporal preferences, TPO offers a robust and efficient solution for advancing temporal reasoning in multimodal tasks. One future direction is scaling the preference data to improve coverage and diversity, thereby enhancing TPO's generalizability. Additionally, while this work focuses on LongVA-7B and LLaVA-Video-7B as representative Video-LMMs, applying TPO to a broader range and larger scale of video-LMMs would provide insights into its adaptability and performance across different architectures.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

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

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*, 2024a.

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning large multimodal models for videos using reinforcement learning from ai feedback. *arXiv preprint arXiv:2402.03746*, 2024b.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale. In *CVPR*, 2025a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024a.

Ruizhe Chen, Zhiting Fan, Tianze Luo, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video temporal grounding via reinforcement learning. *arXiv preprint arXiv:2507.18100*, 2025b.

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*, 2024b.

Shuo Chen, Gang Niu, Chen Gong, Jun Li, Jian Yang, and Masashi Sugiyama. Large-margin contrastive learning with distance polarization regularizer. In *International Conference on Machine Learning*, pp. 1673–1683. PMLR, 2021.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. LongVILA: Scaling long-context visual language models for long videos. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL `https://openreview.net/forum?id=wCXAlfvCy6`.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FZW7Ctyjm3.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024.

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.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *ECCV*, 2024.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. Bimba: Selective-scan compression for long-range video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29096–29107, 2025.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

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

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next: Tackling multi-image, video, and 3d in large multimodal models, June 2024c. URL https://llava-vl.github.io/blog/2024-06-16-llava-next-interleave/.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023a.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

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*, 2024b.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2024c. URL https://arxiv.org/abs/2412.04468.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024d.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23023–23033, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. *arXiv preprint arXiv:2403.08730*, 2024.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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*, pp. 14313–14323, 2024.

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

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv:2410.17434*, 2024.

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.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

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*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6847–6857, 2021.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2025a.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025b.

Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. *arXiv e-prints*, pp. arXiv–2503, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL `https://arxiv.org/abs/2407.15754`.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10714–10726, 2023.

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.

Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126:375–389, 2018.

Yin Song and Chen Wu and Eden Duthie. aws-prototyping/long-llava-qwen2-7b, 2024. URL https://huggingface.co/aws-prototyping/long-llava-qwen2-7b.

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024a. URL https://arxiv.org/abs/2407.12772.

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*, 2024b.

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*, 2024c.

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, April 2024d. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024e. URL https://arxiv.org/abs/2410.02713.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

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*, 2024a.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024b.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

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*, 2024a.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An exploration of video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*, 2024b.

A Reproducibility Statement

We have uploaded the source code and data to the supplementary materials. To ensure reproducibility, we release the full TPO pipeline (scripts and source code), the curated preference dataset (videos, queries, and preference responses/pairs), together with configuration files, prompts, and exact commands to reproduce all tables and figures. By providing these resources, we aim to facilitate the replication of our results and support further advancements in this area of research.

B Use of Large Language Models (LMMs)

In this work, LMMs are used only for polishing the manuscript, with all outputs verified and corrected by humans.

C Appendix overview

This document provides more details of our approach and additional experimental results, organized as follows:

- § D More Implementation Details of TPO.
- § E More Details of the Preference Dataset Curation.
- § F More Ablations on LLaVA-Video
- § G More Examples in the Preference Dataset.
- § H More Qualitative Examples.

D Implementation Details

We conduct Temporal Preference Optimization (TPO) on LongVA (Zhang et al., 2024b) and LLaVA-Video (Zhang et al., 2024e), two state-of-the-art video-LMMs. The two TPO models are trained using 8 Nvidia A100 80GB GPUs, with a batch size of 64. For preference optimization, we set the KL-divergence weight (β) to 0.3 and the supervised fine-tuning (SFT) loss weight (α) to 0.5 for LongVA-TPO and we set the KL-divergence weight (β) to 0.2 and the supervised fine-tuning (SFT) loss weight (α) to 1 for LLaVA-Video-TPO. We employ full fine-tuning for both the multimodal projector and the language model while keeping the visual encoder frozen, using a learning rate of 4×10^{-6} for LongVA-TPO and 3×10^{-7} for LLaVA-Video-TPO. The training is performed on a curated dataset of 10k samples for one epoch for LongVA-TPO and 10k samples for one epoch for LLaVA-Video-TPO. A cosine learning rate scheduler with a warm-up ratio of 0.1 is utilized (Loshchilov & Hutter, 2016). The entire TPO fine-tuning process takes approximately 4 hours on both two models.

For evaluation, we adopt the protocol outlined by LongVA (Zhang et al., 2024b) and LLaVA-Video (Zhang et al., 2024e), leveraging the official lmms-eval repository (Zhang et al., 2024a) to assess our model's performance on three benchmarks. For LongVA-TPO, we set the parameter $max_frames_num = 128$ across all three benchmarks. For LLaVA-Video-TPO, we set the parameter $max_frames_num = 96$ for the Video-MME benchmark and $max_frames_num = 128$ for the rest of the benchmarks.

E Preference Dataset Curation

We manually curated a set of 200 keywords assisted with GPT-4o-mini (Achiam et al., 2023), which were utilized to retrieve 8,000 videos from the internet, forming a diverse and comprehensive dataset. Using this dataset, we further developed 10,000 queries paired with their corresponding preference responses, covering a broad range of tasks. The detailed prompts for preference dataset curation are provided in Fig. 9 and Fig. 10. For LLaVA-Video, we sampled a subset of 10k QA pairs from the LLaVA-Video-178k dataset with the negative responses only curated by incomplete videos.

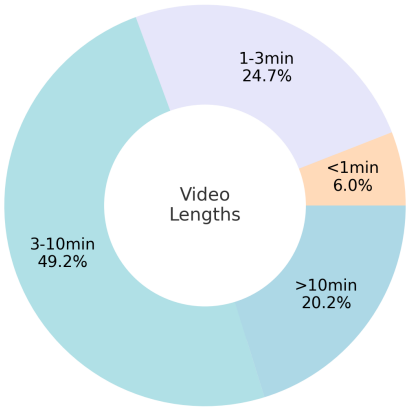


Figure 7: The distribution of lengths for 8K crawled videos.

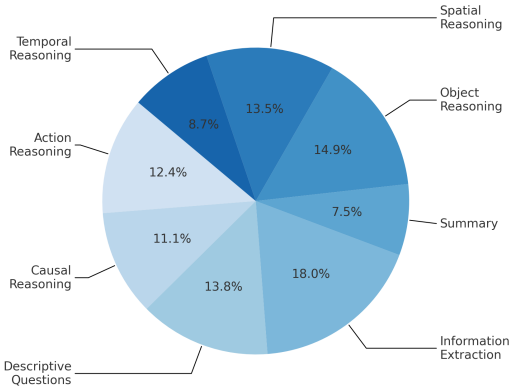


Figure 8: The distribution of question types for 10K curated preference dataset for LongVA-TPO.

The distribution of video lengths in our collected dataset is presented in Fig. 7. The distribution of tasks is illustrated in Fig. 8, encompassing Temporal Reasoning (8.7%), Action Reasoning (12.4%), Causal Reasoning (11.1%), Information Extraction (18.0%), Descriptive Questions (12.8%), Summarization (7.5%), Object Reasoning (14.9%), and Spatial Reasoning (13.5%).

F Ablations on LLaVA-Video

We apply both SFT_{self} and SFT_{LLM} strategies on LLaVA-Video-7B (Zhang et al., 2024e). Notably, all training strategies (including TPO) utilize a subset of the LLaVA-Video SFT dataset. Simply continuing SFT training on a subset does not lead to performance boost. The observed gains from our method highlight that the improvement stems from the design of TPO itself, rather than from the scale of the training data.

Model	LongVideoBench	Video-MME (w/o subs)			
		Short	Medium	Long	Average
LLaVA-Video	58.2	-	-	-	63.3
$+\text{SFT}_{\text{self}}$	57.9	76.1	61.7	53.6	62.8
$+\text{SFT}_{\text{LLM}}$	58.5	76.0	59.9	52.1	62.6
LLaVA-Video-TPO	**60.1**	**76.8**	**64.6**	**55.4**	**65.6**

G Preference Dataset Examples

We provide three additional examples of preference datasets, as illustrated in Fig. 11. For instance, in Example (a), the task involves an OCR-based query aimed at retrieving the quote located beneath a mural. The dis-preferred response incorrectly identifies the relevant frame, failing to locate the quote below the mural and instead referencing another frame containing the phrase “Forward, Warrior.” In contrast, the preferred response accurately identifies the corresponding frame based on the question. This is achieved by leveraging the highly relevant sub-video segment provided to the video-LMM, enabling the correct extraction of both the quote and its attribution.

For Example (b), the task involves summarizing information by identifying the four levels depicted in a pyramid diagram. The dis-preferred response, based on irrelevant video clips, provides incorrect names and an incorrect order for the four levels. In contrast, the preferred response accurately identifies both the correct names and the proper order of the four levels, demonstrating a better understanding of the context and alignment with the video content.

For Example (c), the task involves a high-level descriptive query requiring a summary of the exercise routine depicted in the video. The dis-preferred response, relying only on down-sampled frames, omits significant key information and provides an incomplete summary. In contrast, the preferred

response accurately summarizes the entire exercise routine, offering both detailed and correctly ordered information, thereby demonstrating a comprehensive understanding of the video content.

H QUALITATIVE ANALYSIS EXAMPLES

We provide three additional qualitative analysis examples from the Video-MME dataset (Fu et al., 2024), as illustrated in Fig. 12. Example (a) involves an information extraction and optical character recognition (OCR) task, where the question asks for the total number of measurements involved in chip manufacturing. The original LongVA model failed to accurately locate the relevant frame containing the necessary information, resulting in an incorrect response. In contrast, our LongVA-TPO model, enhanced through temporal preference optimization, successfully identified the pertinent frame within the lengthy input and provided the correct answer to the question.

Example (b) involves a high-level video understanding and information extraction task, where the question asks for the main topic introduced in the video. The original LongVA model failed to capture the overarching theme, instead responding with an unrelated term, "Criminal Trial," mentioned elsewhere in the video. In contrast, our LongVA-TPO model effectively identified the video's central theme and accurately provided the correct topic introduced in the content.

Example (c) involves an object reasoning task, where the question asks what the three curved lines extending from the bottom upward symbolize. The original LongVA model failed to interpret the representation accurately, erroneously stating that the lines represent three stages of the water cycle, which was a hallucination. In contrast, our LongVA-TPO model successfully understood the symbolic meaning of the three curved lines as representing evaporation, providing a correct and detailed response.

```
<Video Caption>

Using the caption of a video, create a question-answer pair that focuses on <Task Prompt>.

Please generate a question tailored to the given caption. If it's inappropriate to generate
    ↪ such question, please output None.
Output format:
Q: <question>
A: <answer>
```

Figure 9: **Detailed prompt for the query generation given the video captions.**

```
<Video Caption>

Question: <Query>
Answer1: <Preferred Answer>
Answer2: <Dis-Preferred Answer>

Task1: You are given a question, the golden answer and related captions. Is answer1 better
    ↪ than answer2?  Please answer with Yes or No or Equally.

Task2: Please check if this question and Answer1 contradicts to any part of the golden
    ↪ caption or this question might have another answer different from the given answer.
    ↪ Please respond with Yes or No.

Task3: Is the Answer1 is correct given the question and golden caption? Please respond with
    ↪ Yes or No.
```

Figure 10: **Detailed prompt for the post-filtering process for the preference data pairs.**

Query: What quote is featured below the mural in the video, and who is it attributed to?

Dis-preferred Response: The quote below the mural is "Forward, Warrior!" and it is attributed to a woman.

Preferred Response: The quote below the mural in the video is "I appeal to all of you to get into this great revolution that's sweeping this nation. Get in and stay in the streets of every city, every village and hamlet of this nation until true freedom comes, until the revolution of 1776 is complete." This quote is attributed to John Lewis.

(a)

Query: What are the four levels represented in the pyramid diagram of the wine tasting guide, and how do they classify wine quality?

Dis-preferred Response: The pyramid diagram in the wine tasting guide represents four levels of wine quality classification: AOC, IGP, Vin de Pays, and Vins. These levels are arranged in a descending order of quality …

Preferred Response: The pyramid diagram in the wine tasting guide represents four levels of wine quality classification: AOC, VDQS, IGP, and Vin. The pyramid structure indicates that each level is a subset of the one above it, with AOC being the highest and Vin being the lowest. …

(b)

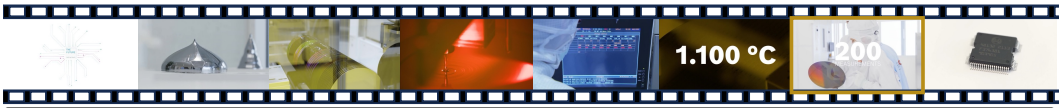
Query: Describe the exercise routine depicted in the video, including the name of the exercises, the number of repetitions, and any additional instructions provided.

Dis-preferred Response: The video shows a man performing a mountain climber exercise, which involves kicking the feet back as the hands touch the floor. He also does a walkout push-up, which involves lifting the body off the ground and then returning to the starting position. The video provides the number of repetitions for each exercise, which are 10 to 12 reps.

Preferred Response: The video shows a man performing a variety of exercises in a gym. He starts by lifting a dumbbell over his head, then proceeds to do mountain climbers, burpees, kettlebell swings, and push-ups. The video also includes a summary of the workout, which includes the following exercises and repetitions: 1. Dumbbell Squat Press (10-12 reps) 2. Mountain Climbers (10-12 reps) 3. Burpees (10-12 reps) 4. Kettlebell Swings (10-12 reps) 5. Walkout Push-ups (10-12 reps). The man is wearing a black shirt and shorts, and he appears to be in good physical shape.

(c)

Figure 11: Examples from the preference dataset.



Query: Based on the video, what is the total number of measurements involved in chip manufacturing?

LongVA: The video shows a sequence of images with the text \"270 measures\" at different stages of chip manufacturing. This suggests that there are 270 measurements involved in the process of chip manufacturing as depicted in the video.

LongVA-TPO: The video states that there are 200 measurements involved in chip manufacturing.

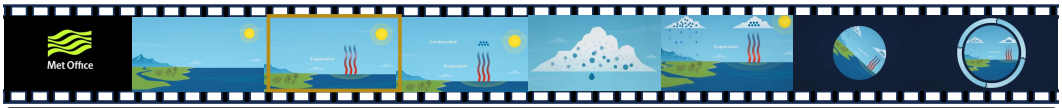
(a)

Query: What topic is introduced in the video?

LongVA: The video introduces the topic of \"Criminal Trial\" with a humorous and animated approach. It features a courtroom setting … The characters are drawn in a cartoonish style, which adds to the comedic effect.

LongVA-TPO: The video introduces the topic of the roles of different people in a courtroom.

(b)



Query: According to the video, what do the three curved lines extending from bottom up symbolize?

LongVA: The three curved lines extending from bottom up in the image are likely representing the three main stages of the water cycle: evaporation, condensation, and precipitation. These stages are typically depicted in educational materials to illustrate how water changes state and moves through the environment.

LongVA-TPO: The three curved lines extending from bottom up in the image symbolize the process of evaporation. They represent the upward movement of water vapor as it transitions from a liquid state to a gaseous state, rising into the atmosphere. This is a key part of the water cycle and demonstrates the concept visually.

(c)

Figure 12: Additional qualitative examples from Video-MME (Fu et al., 2024).