TSPO: Temporal Sampling Policy Optimization for Long-form Video Language Understanding

Canhui Tang[1,2*], Zifan Han[1,2*], Hongbo Sun[2], Sanping Zhou[1†], Xuchong Zhang[1], Xin Wei[2], Ye Yuan[2], Huayu Zhang[2], Jinglin Xu[3], Hao Sun[2‡]

[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]University of Science and Technology Beijing

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated significant progress in vision-language tasks, yet they still face challenges when processing long-duration video inputs. The limitation arises from MLLMs' context limit and training costs, necessitating sparse frame sampling before feeding videos into MLLMs. However, building a trainable sampling method remains challenging due to the unsupervised and non-differentiable nature of sparse frame sampling in Video-MLLMs. To address these problems, we propose **T**emporal **S**ampling **P**olicy **O**ptimization (**TSPO**), advancing MLLMs' long-form video-language understanding via reinforcement learning. Specifically, we first propose a trainable event-aware temporal agent, which captures event-query correlation for performing probabilistic keyframe selection. Then, we propose the TSPO reinforcement learning paradigm, which models keyframe selection and language generation as a joint decision-making process, enabling end-to-end group relative optimization for the temporal sampling policy. Furthermore, we propose a dual-style long video training data construction pipeline, balancing comprehensive temporal understanding and key segment localization. Finally, we incorporate rule-based answering accuracy and temporal locating reward mechanisms to optimize the temporal sampling policy. Comprehensive experiments show that our TSPO achieves state-of-the-art performance across multiple long video understanding benchmarks, and shows transferable ability across different cutting-edge Video-MLLMs.

Code — https://github.com/Hui-design/TSPO

Introduction

Multimodal Large Language Models (MLLMs) have achieved significant progress in various vision-language tasks, such as image captioning, visual question answering, OCR, etc. They typically extract visual information as visual tokens into the Large Language Models (LLMs) for open-world understanding. As a natural extension, video-based MLLMs (Video-MLLMs) (Zhang et al. 2024d; Shen et al. 2024; Bai et al. 2025) have attracted great attention, where videos contain more complex temporal and visual information, bringing more significant challenges.

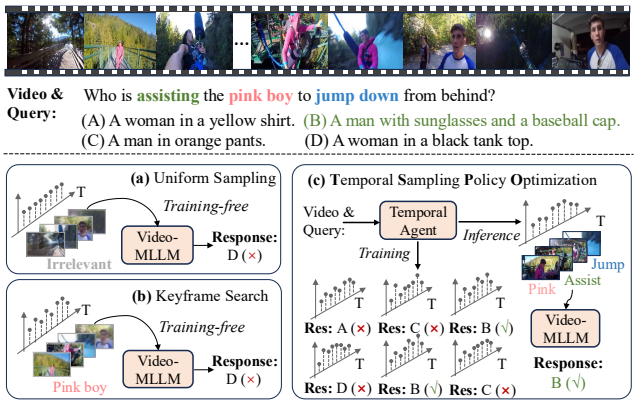

Figure 1: Illustrations of different frame sampling methods: Training-free uniform sampling (a) and keyframe search (b) select unsatisfactory frames, while our method (c) explores and optimizes the temporal sampling policy that leads to the correct answer in an end-to-end training manner.

Existing MLLMs are compelled to employ sparse frame sampling when dealing with videos (Zhang et al. 2024d; Shen et al. 2024; Kim et al. 2024; Xu et al. 2024b; Liu et al. 2024c). The core challenge mainly lies in determining the optimal frame sampling strategy that maximizes MLLMs' video comprehension accuracy while minimizing computational overhead. Most existing Video-MLLM approaches, such as LLaVA-Video (Zhang et al. 2024d) and Qwen2.5VL (Bai et al. 2025), simply perform uniform frame sampling, which often misses key information that is relevant to queries, as shown in Fig. 1. Recently, some studies (Hu et al. 2025a; Shen et al. 2024; Tang et al. 2025) focus on exploring training-free keyframe extraction approaches. For instance, LongVU (Shen et al. 2024) identifies cross-frame distinct frames by leveraging pre-trained feature extractors such as DINOv2-1B (Oquab et al. 2023). CoS (Hu

*These authors contributed equally.
†Corresponding authors.
‡Corresponding authors.

et al. 2025a) employs LLaVA-1.5-13B (Liu et al. 2024a) to filter query-relevant frames for inputting into Video-MLLM, which incurs significant computational costs. Without training optimization, training-free methods are limited by the cross-modal event understanding capabilities of pre-trained keyframe selectors and may incur more computation during inference. These limitations lead to a question: *Can we develop a trainable sparse frame sampling approach for reliable and efficient long-video language understanding?*

However, there exist two fundamental challenges to obtaining a trainable temporal sampling approach for Video-MLLMs: **(1) Unsupervised nature**: frame-level annotations are generally unavailable in general video understanding training (Zhang et al. 2024d; Bai et al. 2025), resulting in a lack of precise localization guidance. **(2) Non-differentiability**: frame sampling is a discrete subset selection problem, where the output consists of frame indices rather than continuous variables, making it difficult to optimize via backpropagation in Supervised Fine-Tuning (SFT).

Based on the above analyses and inspired by the progress of Deepseek-R1 (DeepSeek-AI et al. 2025; Shao et al. 2024) in enhancing MLLM reasoning through Group Relative Policy Optimization (GRPO), we propose Temporal Sampling Policy Optimization (TSPO) to explore and optimize keyframe selection strategy for Video-MLLMs. It novelly models keyframe selection and language generation as a joint decision-making process, performing end-to-end GRPO optimization of the temporal agent through rule-based rewards. Specifically, a trainable temporal sampler is first modeled as a decision agent to capture event-query correlation for keyframe probability estimation, which also maintains structural simplicity instead of using other heavy MLLMs like (Hu et al. 2025a,b). Furthermore, for TSPO's training, we propose a long video training data construction pipeline with *comprehensive temporal data* for general video understanding and *video Needle-in-a-Haystack* data for long-range temporal localization. In the reinforcement learning-based temporal sampling policy optimization, we establish efficient rule-based answering accuracy and coarse-level temporal locating reward mechanisms that optimize the temporal agent to maximize the expected reward by choosing critical frames for queries adaptively.

Extensive experiments demonstrate the effectiveness and strong generalization of our TSPO method, achieving average performance gains of **4.3%** on LLaVA-Video and **6.1%** on Qwen2.5-VL. Our contributions are as follows:

- We propose the Temporal Sampling Policy Optimization algorithm, which models keyframe selection and language generation as a joint decision-making process, performing end-to-end group relative optimization for the temporal sampling policy. This effectively tackles the unsupervised and non-differentiable challenge of sparse frame sampling in Video-MLLMs.
- We propose a TSPO-targeted training data construction pipeline with comprehensive temporal data and Video Needle-in-a-Haystack data, incorporating the establishment of rule-based answering accuracy and temporal locating reward mechanisms.
- Our TSPO achieves state-of-the-art performance across multiple general long video understanding benchmarks, and shows strong transferable ability across different cutting-edge Video-MLLMs.

Related Work

MLLMs for Long Video Understanding

Multimodal Large Language Models (MLLMs) have demonstrated significant progress in vision-language tasks, yet they still face challenges when processing long-duration videos with extremely long context. Previous works (Xu et al. 2024a; Xu, Yin, and Peng 2025; Xu et al. 2025a,b, 2024b; Liu et al. 2024b; Lin et al. 2023; Cheng et al. 2024; Zhang et al. 2024d,b) often employ uniform frame sampling or perform token compression to reduce the length of the context. LLaVA-Video (Zhang et al. 2024d) and SlowFast-LLaVA (Xu et al. 2024b) utilize spatial and temporal pooling techniques to decrease the number of tokens. Beyond uniform sampling, recent works are exploring keyframe search methods (Ye et al. 2025; Guo et al. 2025; Wang et al. 2023; Shen et al. 2024; Hu et al. 2025a). For instance, LongVU (Shen et al. 2024) identifies cross-frame distinct frames by leveraging robust feature extractors such as DINOv2 (Oquab et al. 2023) and further discards tokens that exhibit minimal feature differences between the current frame and those of the previous frames. More recently, Chain-of-Shot (Hu et al. 2025a) leverages off-the-shelf multimodal large models, such as LLaVA-1.5 (Liu et al. 2024a), to select task-relevant shots and task-irrelevant shots. However, due to the inherently non-differentiable and unsupervised nature of keyframe sampling in general video understanding tasks, these methods opt for a training-free approach, which cannot be further optimized. In this paper, we propose a learnable temporal agent and present the TSPO algorithm to optimize keyframe sampling.

Reinforcement Learning for MLLMs

Reinforcement Learning (RL) is often used in post-training of LLMs to align with human preferences. To enhance multi-modal capabilities (Wang et al. 2024b; Zhang et al. 2024c; DeepSeek-AI et al. 2025). RLHF(Ouyang et al. 2022) utilizes human feedback to train a reward model, treating the LLM as a policy network and optimizing it using PPO (Schulman et al. 2017). DPO (Rafailov et al. 2024) simplifies RLHF by directly constructing preference data without requiring a reward model. LLaVA-Hound-DPO (Zhang et al. 2024c) and TPO (Li et al. 2025) construct temporal preference data and optimize the LLM within the DPO framework. More recently, Deepseek-R1 (DeepSeek-AI et al. 2025) has garnered significant attention by using the GRPO (Shao et al. 2024) algorithm to achieve robust reasoning capabilities. However, these reinforcement learning approaches focus solely on optimizing the reasoning abilities of LLMs. In contrast, our TSPO takes a novel perspective by modeling the discrete keyframe selection and language generation as a unified decision-making process, directly addressing the most challenging issue in current long video understanding: the extremely long context problem.

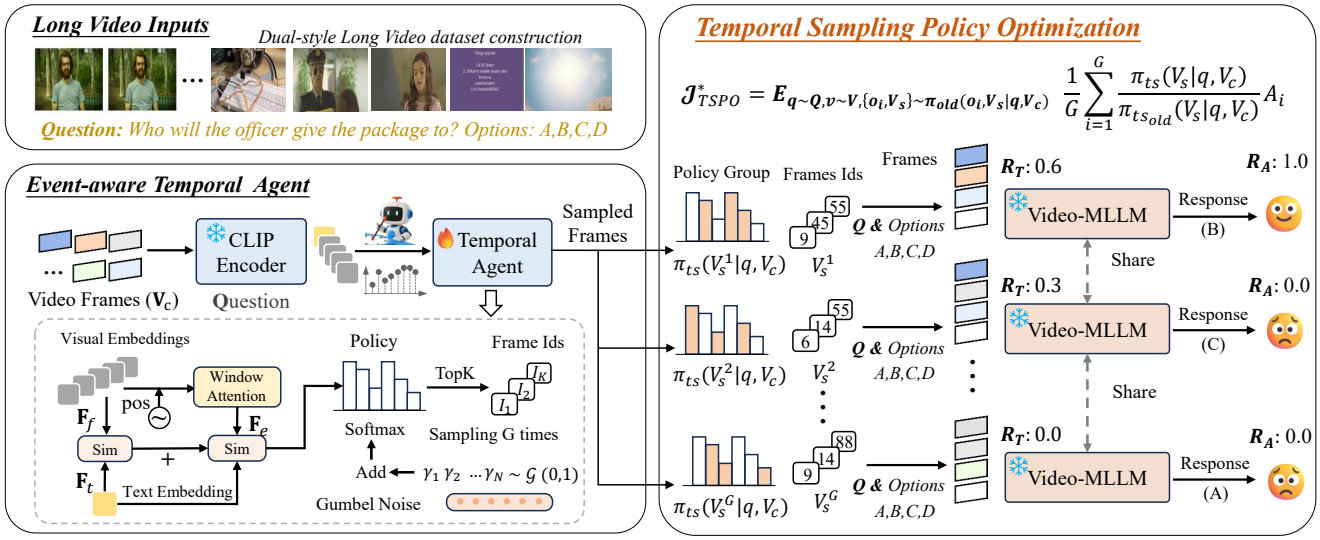

Figure 2: The overview of our TSPO framework. The training pipeline takes long videos as inputs, first employing a temporal agent to sample G keyframe combinations (only one during inference), then optimizing the sampling policy through our temporal sampling policy optimization algorithm with Temporal localization reward R_T and Answering Accuracy reward R_A.

Method

As shown in Fig. 2, we propose Temporal Sampling Policy Optimization (TSPO) to advance MLLMs' long-form video-language understanding via reinforcement learning. First, we model the temporal sampling policy for Video-MLLM by integrating discrete frame sampling into the language model's decision-making process and establishing an event-aware temporal agent for probabilistic keyframe selection. Second, we propose an RL-based temporal optimization approach for Video-MLLMs. Finally, we present our TSPO training dataset construction pipeline and reward designs.

Modeling Temporal Sampling Video-MLLM

Previous works based on supervised fine-tuning or reinforcement learning focus solely on MLLM optimization, while overlooking the optimization for frame selection. Unlike training-free uniform frame sampling or keyframe search, as shown in Fig. 3, we aim to explore RL-based optimization schema by modeling discrete keyframe selection and language generation as a joint decision-making process.

Vanilla Video-MLLM Policy. Video-MLLMs are typically composed of three core components: a visual encoder, a multimodal projector, and a large language model (LLM) π_l. Video-MLLMs take a video $\mathbf{v}$ and a text query $\mathbf{q}$ as inputs. Due to the LLM context limit, the video is first processed into sparse frames $\mathbf{V}_s$ by uniform sampling or training-free selectors. Video-MLLMs encode them into visual tokens and text tokens, respectively. These multimodal tokens are then concatenated and processed by the LLM through autoregressive generation to produce the final textual response. Then Video-MLLM models the likelihood of generating a language response output $\mathbf{o}$ as follows:

$$\pi_l(\mathbf{o} \mid \mathbf{q}, \mathbf{V}_s) = \prod_{i=1}^{n} \pi_l(o_i \mid o_{<i}, \mathbf{q}, \mathbf{V}_s). \quad (1)$$

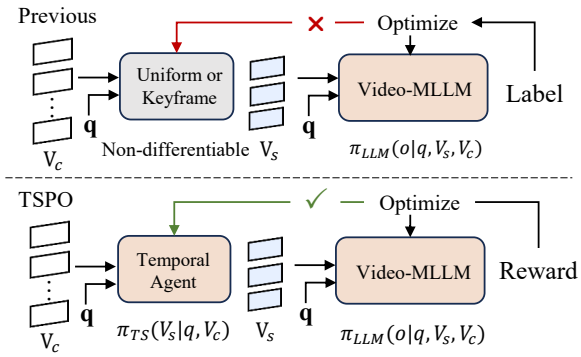

Figure 3: Comparison between our TSPO and previous Video-MLLM optimization methods. We model keyframe selection and language generation as a joint decision-making process for end-to-end optimization of the temporal agent.

Temporal Sampling Video-MLLM Policy. Our TSPO first integrates frame sampling into the decision-making process of Video-MLLMs. Considering computing cost, uniform sampling is first applied to obtain T_c candidate frames from a T-frame video. Then, adaptive keyframe selection is conducted based on textual query $\mathbf{q} \in \mathbf{Q}$ and frame-level visual features, yielding a keyframe combination with T_s-frames. The temporal sampling policy is formulated as $\pi(\mathbf{y}, \mathbf{V_s} \mid \mathbf{q}, \mathbf{V}_c)$. Following the conditional probability rule and the chain rule, this policy can be expressed as:

$$\pi(\mathbf{o}, \mathbf{V_s} \mid \mathbf{q}, \mathbf{V}_c) = \pi_l(\mathbf{o} \mid \mathbf{q}, \mathbf{V}_s, \mathbf{V}_c) \cdot \pi_{ts}(\mathbf{V}_s \mid \mathbf{q}, \mathbf{V}_c), \quad (2)$$

where $\mathbf{V}_c$ is the candidate video frames, and $\mathbf{V}_s$ denotes the selected keyframes.

Event-aware Temporal Agent

To model the $\pi_{ts}(\mathbf{V}_s \mid \mathbf{q}, \mathbf{V}_c)$ policy, we first propose a trainable event-aware temporal agent, which captures event-query correlation and performs probabilistic keyframe selection from an RL perspective.

As shown in Fig. 2, the temporal agent takes CLIP (Radford et al. 2021) frame-level visual features $\mathbf{F}_f \in \mathbb{R}^{T_c \times D}$ and text features $\mathbf{F}_t \in \mathbb{R}^{1 \times D}$ as the inputs, where T_c denotes the candidate frame number and D denotes the feature dimension. The visual features are then enhanced with event perception capabilities through local window attention (Pu et al. 2024). Then the attention is restricted to a local window of length w centered at the current frame, augmented with sinusoidal positional embeddings (Vaswani et al. 2017), and projected by an MLP to learn intra-event dependencies and temporal awareness. This leads to the refined event representation $\mathbf{F}_e \in \mathbb{R}^{T_c \times D}$, and the cosine similarity $\mathrm{Sim}_{event}(\mathbf{F}_e, \mathbf{F}_t)$ is then computed to capture event-text alignment. To strengthen temporal localization robustness, frame-level similarity $\mathrm{Sim}_{frame}(\mathbf{F}_f, \mathbf{F}_t)$ between visual features F_v and text features F_t is concurrently calculated. The final cross-modal similarity $S \in \mathbb{R}^{T_c}$ is derived through the fusion of the two scores:

$$S = \mathrm{Sim}_{event}(\mathbf{F}_e, \mathbf{F}_t) + \mathrm{Sim}_{frame}(\mathbf{F}_f, \mathbf{F}_t). \tag{3}$$

In the reinforcement learning framework, the **agent state** is defined by the input long video V and the text instruction Q. The **agent action** corresponds to keyframe selection, outputting the selected indices $\mathcal{I} = \{i_1, i_2, .., i_{T_s}\}$ and the corresponding probability $\mathcal{P} = \{p_1, p_2, .., p_{T_s}\}$. To enable diverse action generation for RL exploration (Wei et al. 2023; Cui et al. 2023), our method employs the Gumbel-Softmax (Jang, Gu, and Poole 2016) operator:

$$\mathcal{P}, \mathcal{I} = \mathrm{TopK}\left(\mathrm{Softmax}(S/\tau + \gamma)\right), \gamma \sim \mathrm{Gumbel}(0, 1), \tag{4}$$

where τ is the temperature parameter. The generation process injects Gumbel (0,1) noise into cross-model scores, then selects the Top-T_s query-relevant frames with their corresponding probabilities. The probability is expressed by:

$$\pi_{ts}(\mathbf{V}_s \mid \mathbf{q}, \mathbf{V}_c) = \prod_{i=1}^{T_s} \mathcal{P}_i(\mathbf{V}_c, \mathbf{q}). \tag{5}$$

Temperature annealing. A high temperature is used early in training to encourage exploration, and it is gradually reduced to a low temperature to converge to key segments. Then, the selected frames indices $\mathcal{I}$ are used to obtain the sparse frames, which serve as inputs to the Video-MLLMs.

Temporal Sampling Policy Optimization

In this part, we present a novel multimodal temporal sampling policy optimization algorithm that enables end-to-end group relative optimization of keyframe selection.

Vanilla Group Relative Policy Optimization (GRPO). Deepseek-R1 (DeepSeek-AI et al. 2025) proposes the GRPO (Shao et al. 2024) algorithm, which foregoes the critic model and estimates the baseline from rule-based rewards instead. Specifically, for each question $\mathbf{q}$ from the training set $\mathbf{Q}$, a group of language outputs is sampled $\{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ by maximizing:

$$\mathcal{J}_{grpo}(\theta) = \mathbb{E}_{\mathbf{q}\sim\mathbf{Q}, \{\mathbf{o}_i\}\sim\pi_{\theta_{old}}(\mathbf{O}|\mathbf{q})} \frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(\mathbf{o}_i|\mathbf{q})}{\pi_{\theta_{old}}(\mathbf{o}_i|\mathbf{q})}A_i - \beta \cdot \mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \tag{6}$$

where $\mathbb{D}_{KL}$ is the unbiased estimator (Schulman 2020) for KL divergence and β is a hyper-parameter to balance the weights. π_{ref} is the reference model, typically a LLM that has undergone large-scale Supervised Fine-Tuning (SFT). A_i is the relative advantage, computed using a group of rewards $\{r_1, r_2, \ldots, r_G\}$ corresponding to the group outputs (DeepSeek-AI et al. 2025).

Temporal Sampling Policy Optimization (TSPO). As shown in Fig. 3, our TSPO models keyframe selection and language generation as a joint decision-making process, enabling end-to-end GRPO optimization through language supervision. In detail, the temporal agent and Video-MLLMs are treated as a policy pool capable of making probabilistic estimations for frame selection and response generation, as shown in Eq. (2). Then, the decision process is supervised by maximizing the expected reward of actions. Therefore, the objective can be reformulated as follows:

$$\mathcal{J}_{tspo}(\theta) = \mathbb{E}_{\mathbf{q}\sim\mathbf{Q}, \mathbf{v}\sim\mathbf{V}, \{\mathbf{o}_i, \mathbf{V}_s\}\sim\pi_{old}(\mathbf{O} \mid \mathbf{q}, \mathbf{V}_c)} \frac{1}{G}\sum_{i=1}^{G}\frac{\pi_l(\mathbf{o}_i \mid \mathbf{q}, \mathbf{V}_s, \mathbf{V}_c) \cdot \pi_{ts}(\mathbf{V}_s \mid \mathbf{q}, \mathbf{V}_c)}{\pi_{l_{old}}(\mathbf{o}_i \mid \mathbf{q}, \mathbf{V}_s, \mathbf{V}_c) \cdot \pi_{ts_{old}}(\mathbf{V}_s \mid \mathbf{q}, \mathbf{V}_c)}A_i - \beta \cdot \mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right). \tag{7}$$

Considering the extremely long context problem is more significant for the current long video understanding model, we maintain focus on optimizing the Temporal Sampler while preserving the strong prior of language generation capabilities. We employ a pre-trained MLLM and keep it frozen, thereby ensuring that:

$$\pi_l(\mathbf{o}_i \mid \mathbf{q}, \mathbf{V}_s, \mathbf{V}_c) \,/\, \pi_{l_{old}}(\mathbf{o}_i \mid \mathbf{q}, \mathbf{V}_s, \mathbf{V}_c) = 1. \tag{8}$$

Notably, the MLLM has been SFT-trained on LLaVA-Video-178K (our source dataset) with uniform 32 frames and thus can answer questions well when correct keyframes are selected. Therefore, our TSPO objective can be simplified to optimize only the temporal agent as follows:

$$\mathcal{J}^*_{tspo}(\theta) = \mathbb{E}_{\mathbf{q}\sim\mathbf{Q}, \mathbf{v}\sim\mathbf{V}, \{\mathbf{o}_i, \mathbf{V}_s\}\sim\pi_{old}(\mathbf{O} \mid \mathbf{q}, \mathbf{V}_c)} \frac{1}{G}\sum_{i=1}^{G}\frac{\pi_{ts}(\mathbf{V}_s \mid \mathbf{q}, \mathbf{V}_c)}{\pi_{ts_{old}}(\mathbf{V}_s \mid \mathbf{q}, \mathbf{V}_c)}A_i\,, \tag{9}$$

where the advantage A_i is computed through an efficient rule-based reward mechanism.

TSPO Training Dataset and Reward Design

To drive TSPO training, we introduce *comprehensive temporal data* for general video understanding and *video Needle-in-a-Haystack data* for long-range temporal localization, as shown in Fig. 4. The training incorporates question answering accuracy and temporal localization rewards.

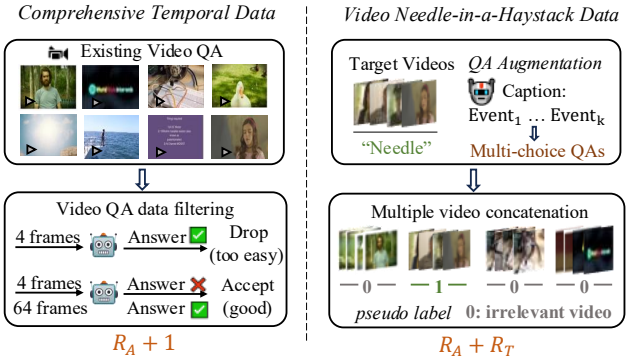

Figure 4: Our proposed TSPO-targeted long video training data construction pipeline.

(1) Comprehensive Temporal Data. Thanks to our TSPO's capability for end-to-end language-guided optimization without frame-level annotations, we can reuse existing video QA datasets with little effort. We collect video multiple-choice QA data longer than 1 minute from LLaVA-Video-178K (Zhang et al. 2024d) (Video max length: 3 minutes). Furthermore, to increase data quality, we filter items that are answerable from 4 uniform frames (too easy) or unsolvable even when sampling 64 frames from a 1-to-3-minute video (too hard). The remaining data requires sampling multiple keyframes, featuring comprehensive temporal dependency. **(2) Video Needle-in-a-Haystack.** The Video-MLLMs community still lacks high-quality long-video QA datasets. For instance, the prominent LLaVA-Video-178K dataset contains videos no longer than 3 minutes. Inspired by the "Needle-in-a-Haystack" designed for evaluation (Zhang et al. 2024b), we propose a long video training data construction pipeline. We sample videos from LLaVA-Video-178K as target videos, applying QA augmentation since some original training questions are too generic to localize segments in spliced videos. Using Qwen2.5-VL (Bai et al. 2025), we generate detailed event descriptions for target videos, reformatted into multiple-choice questions. Finally, the target videos are concatenated and shuffled with irrelevant videos at the segment level to form long training videos (10∼60 minutes).

The dual pipelines yield **TSPO-10K**, a high-quality long video dataset comprising 10,000 samples specifically optimized for temporal sampling policy training.

Dual Reward Designs. First, we follow an intuition that Video-MLLMs can only give correct answers if the temporal agent samples the correct keyframes. Thanks to our TSPO modeling, we can use the language response accuracy derived from multiple-choice training data to supervise the temporal agent. The accuracy reward is defined as:

$$R_A = \mathbf{1}(y = \overline{y}), \tag{10}$$

where $\mathbf{1}$ is the indicator function, y is the predicted option, and $\overline{y}$ is the ground-truth option.

For the video Needle-in-a-Haystack task, we leverage the pseudo-labels from our video synthesis pipeline. We quantify the localization precision by computing the ratio of correctly sampled frames to total frames:

$$R_T = T_t \,/\, T_a, \tag{11}$$

where T_t is the count of frames residing in the target video, and T_a is the total sampled frames. For data from "Comprehensive temporal", the total reward is $R_A + 1$, while for data from "Needle-in-a-Haystack", the reward is $R_A + R_T$.

Experiments

Experimental Settings

Evaluation Benchmarks. To evaluate the effectiveness of the proposed TSPO, we conduct experiments on four widely used benchmarks in long-form video understanding.

- **LongVideoBench** (Wu et al. 2024). We evaluate the validation set with 1,337 videos (avg. 12min), following standard academic protocols (Zhang et al. 2024d).
- **MLVU** (Zhou et al. 2024). The video length ranges from 3 minutes to 2 hours. We evaluate on the "M-Avg" portion of the "Dev" split, following (Zhang et al. 2024d).
- **Video-MME** (w/o sub) (Fu et al. 2024) comprises 900 videos with variable durations: short ($<$ 2 min), medium (4∼15 min), and long (30∼60min), containing 2700 QA.
- **LVBench** (Wang et al. 2024c) is an extremely long video benchmark, with an average video length of 4,101 seconds—4 times longer than VideoMME.

Implementation Details. The model was trained on 8 NVIDIA A800 80GB GPUs with a single epoch, using a learning rate of 5×10^{-4} and a batch size of 1. The temporal agent is built upon a frozen CLIP-Large model (400M parameters) and incorporates **only 3.5M learnable** parameters. For the Video-MLLM that guides TSPO training, we adopt LLaVA-Video (Zhang et al. 2024d) with its parameters kept frozen. The number of candidate frames (T_c) is set to 1 FPS, while the selected frame count (T_s) is set to 64 during inference and 16 during training. The window size w is set to 12, τ is set at 0.025 and annealed to 0.01. To ensure reproducibility during inference, deterministic predictions were enforced by removing the Gumbel noise.

Comparison to the State-of-the-art. As shown in Tab. 1, our LLaVA-Video-7B*+TSPO achieves state-of-the-art performance across four general long video benchmarks. Compared to LLaVA-Video-7B*, we improve by **+5.0%** on LongVideoBench, **+6.0%** on MLVU, **+5.1%** on LVBench, and 1.1% on VideoMME. Compared to Qwen2.5VL*, we improve by **+4.9%** on LongVideoBench, **+11.2%** on MLVU, and 1.8% on VideoMME. The modest improvement on VideoMME can be attributed to its emphasis on holistic video comprehension rather than localizations on specific keyframes. Our consistent outperformance over state-of-the-art methods further validates the superiority of our approach in extracting key information from long videos.

Selector Parameters. Experimental comparisons show that our lightweight selector (Temporal Agent) achieves both parameter efficiency and superior model performance. Compared to LongVU's (Shen et al. 2024) 1B-parameter DINOV2 (Oquab et al. 2023) selector, our method achieves

Model	Frames	LLM Size	Selector	LongVideoBench	MLVU	Video-MME		LVBench
				Val	Dev	Long	Average	-
GPT-4o (Hurst et al. 2024)	-	-	Uniform	66.7	64.6	65.3	71.9	-
GPT-4V (OpenAI 2023)	-	-	Uniform	61.3	49.2	53.5	59.9	-
Gemini-1.5-Flash (Team et al. 2023)	-	-	Uniform	61.6	-	61.1	70.3	-
Gemini-1.5-Pro (Team et al. 2023)	-	-	Uniform	64	-	67.4	75.0	33.1
Video-LLaVA (Lin et al. 2023)	8	7B	Uniform	-	36.2	-	39.9	-
Oryx-1.5 (Liu et al. 2025)	64	7B	Uniform	56.3	-	51.2	58.8	-
LLaVA-Onevision (Li et al. 2024)	32	7B	Uniform	56.4	64.7	46.7	58.2	-
NVILA (Liu et al. 2024d)	1024	7B	Uniform	57.7	70.1	54.8	64.2	-
Apollo (Zohar et al. 2024)	2FPS	7B	Uniform	58.5	68.7	-	61.3	-
mPLUG-Owl3 (Ye et al. 2024)	128	7B	Uniform	59.7	70.0	50.1	59.3	43.5
LongVU (Shen et al. 2024)	1FPS	7B	DINOv2-1B	-	65.4	-	60.6	-
MLLM-VFS (Hu et al. 2025b)	32	7B	MLLM-1.5B	57.0	-	51.9	58.7	-
LLaVA-Video-7B (Zhang et al. 2024d)	64	7B	Uniform	58.2	70.8	-	63.3	-
LLaVA-Video-7B+TPO (Li et al. 2025)	64	7B	Uniform	60.1	71.1	55.4	65.6	-
LLaVA-Video-7B+CoS (Hu et al. 2025a)	64	7B	MLLM-13B	58.9	71.4	53.8	64.4	-
LLaVA-Video-7B+AKS (Tang et al. 2025)	64	7B	BLIP-0.5B	62.7	-	54.0	65.3	-
LLaVA-Video-7B* (Zhang et al. 2024d)	64	7B	Uniform	58.9	70.3	53.6	64.4	40.2
LLaVA-Video-7B*+TSPO	64	7B	TSPO-0.4B	63.9	**76.3**	54.7	**65.5**	45.3
Qwen2.5VL* (Bai et al. 2025)	64	7B	Uniform	59.0	65.1	53.3	63.7	38.3
Qwen2.5VL* + TSPO	64	7B	TSPO-0.4B	**64.2**	74.3	**56.4**	**65.5**	**46.4**

Table 1: Comparison results on four widely recognized long video understanding benchmarks, where our method achieves state-of-the-art performances with significant accuracy gain. "*" denotes our reproduced results under 64 frames. The first three benchmarks are evaluated using lmms-eval (Zhang et al. 2024a), and LVBench is tested using its own evaluation protocol.

Model	Param	LongVideoBench	MLVU
LLaVA-Video	7B	58.9	70.3
LLaVA-Video+TSPO	7B	**63.9** $_{5.0\uparrow}$	**76.3** $_{6.0\uparrow}$
LLaVA-Video	72B	62.4	74.4
LLaVA-Video+TSPO	72B	**66.0** $_{3.6\uparrow}$	**77.3** $_{2.9\uparrow}$
Qwen2VL	7B	55.4	64.0
Qwen2VL+TSPO	7B	**59.5** $_{4.1\uparrow}$	**71.0** $_{7.0\uparrow}$
Qwen2.5VL	7B	59.0	65.1
Qwen2.5VL+TSPO	7B	**64.2** $_{5.2\uparrow}$	**74.3** $_{9.2\uparrow}$

Table 2: Performance of transferring TSPO from LLaVA-Video to other Video-MLLMs without extra training, where the sampled frame number is set to **64** consistently.

a 4.9% absolute accuracy improvement on VideoMME and 10.9% on MLVU. Against Chain-of-Shot (Hu et al. 2025a)'s 13B-scale MLLM selector, our solution outperforms by 6.0% on LongVideoBench and 4.9% on MLVU.

Method	Frames	Data	Performance
LLaVA-Onevision+FrameVOYA.	16	12.5K	- / 57.5
LLaVA-Onevision**+TSPO**	16	**10K**	- / **58.7**
Qwen2VL+MLLM-VFS	32	1.5M	57.0 / 58.7
Qwen2VL**+TSPO**	32	**10K**	**58.6 / 59.6**

Table 3: Comparison with recent keyframe training methods under the same settings. The performance is evaluated on LongVideoBench and VideoMME.

Train Data	$\mathbf{R}_A$	$\mathbf{R}_T$	Performance
None	-	-	58.9 / 64.4
Comprehensive Temporal	✓		62.8 / 65.5
Needle-in-a-Haystack		✓	63.4 / 64.6
Needle-in-a-Haystack	✓	✓	63.7 / 64.9
Comprehensive Temporal + Needle.	✓		63.8 / 65.0
Comprehensive Temporal + Needle.	✓	✓	**63.9 / 65.5**

Table 4: Ablation of data curation and reward schemes.

Ablation Study

Transferring TSPO to Other Video-MLLMs. Although our method is developed based on LLaVA-Video as the Video-MLLM backbone, we explore an efficient "one-model for all" paradigm (Liu et al. 2023; Cheng et al. 2025) by transferring our learned temporal agent from LLaVA-Video-7B to other Video-MLLMs **without extra training**, including Qwen2VL (Wang et al. 2024a) / Qwen2.5VL-7B (Bai et al. 2025), and also extending it to LLaVA-Video-72B. As shown in Tab. 2, our method demonstrates notable generalization capability: on LongVideoBench, it achieves an average **4.5%** improvement; On the MLVU dataset, the average improvement is **6.3**%, with Qwen2.5VL achieving a notably higher gain of **9.2**%. This cross-architecture performance verifies the generalizability of our approach.

Comparison with Keyframe Training Method. Both FrameVOYAGER (Yu et al. 2024) and MLLM-VFS (Hu et al. 2025b) are recent training-based methods that require offline keyframe ranking or labeling to supervise the

Method	Data	E2E training	Performance
LLaVA-Video	-	×	58.9 / 64.4
LLaVA-Video+SFT*	30K	✓	62.8 / 64.8
LLaVA-Video+TSPO	10K	✓	**63.9 / 65.5**

Table 5: Comparison results of SFT* and TSPO training.

Method	Frames	Token	Frame Time	LLM Time	Perform.
LLaVA-Video	128→64	13440	0	2.7	58.2 / 63.3
LLaVA-Video + CoS	128→64	13440	28.4	2.7	58.9 / 64.4
LLaVA-Video + TSPO	128→64	13440	1.2	2.7	**60.6 / 65.3**
LLaVA-Video + TSPO	128→32	6720	1.1	1.3	59.6 / 64.8

Table 6: Comparison results of inference efficiency.

keyframe selector. Compared with them, our TSPO offers two advantages: 1) RL modeling: we jointly model the keyframe selection and language generation from an RL perspective, enabling end-to-end optimization with off-the-shelf video QA data, without requiring additional frame annotations like MLLM-VFS. 2) Superior sampling strategy: FrameVOYAGER requires random pre-processing sampling from frame combinations, while our approach is on-policy, dynamically sampling based on the current policy, which progressively refines the selection strategy. For a fair comparison, since neither method has been open-source, we use their reported results from their papers and adapt our TSPO to the same settings. As shown in Tab. 3, TSPO achieves enhanced performance despite using less training data.

Ablation of Training Data. Tab. 4 ablates our training data curation and reward mechanisms. First, directly using "Comprehensive Temporal" data improves performance by 3.9% on LongVideoBench and 1.1% on VideoMME. Next, the "Needle-in-a-Haystack" pipeline boosts LongVideoBench results yet degrades VideoMME performance, as LongVideoBench focuses on long-range temporal localization while VideoMME emphasizes general comprehension. Combining dual-style data achieves the best performance. For the ablation of rewards, using only the accuracy reward guides the temporal agent to select frames that yield correct answers, yet the supervision remains indirect. The temporal reward helps locate relevant clips with coarse labels, yet it may still include irrelevant frames. Combining both rewards enables the model to effectively locate the most relevant frames, leading to correct MLLM answers.

Exploring SFT* for Keyframe. We investigate the possibility of end-to-end (E2E) optimization for the keyframe selector through SFT and compare it with TSPO. We utilize the Gumbel-Softmax technique (Wei et al. 2023), which enables the selected vision token to have gradients (Liang et al. 2024) and can be E2E trained. For SFT* training, we randomly select 30K samples from LLaVA-Video-178K while keeping the MLLM parameters frozen. As shown in Tab. 5, our experimental results demonstrate that TSPO consistently outperforms SFT*, which indicates TSPO's advantages, including its capacity to explore diverse sampling strategies and utilize more direct reward signals.

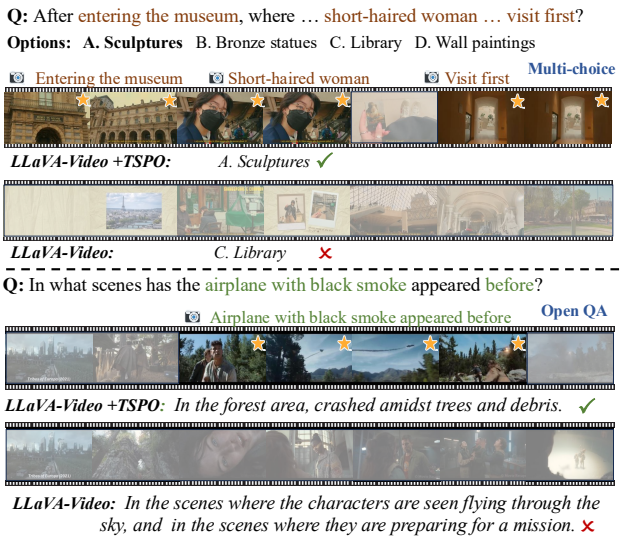

Figure 5: Visualization comparisons of sampled frames and corresponding responses between ours and LLaVA-Video.

Inference Efficiency. In this study, we fix the candidate frame number V_c to 128 (1FPS in our main setting), which avoids the effect of varying video lengths. As shown in Tab. 6, our efficiency can be demonstrated from the following aspects: 1) with the same 64 frames, ours achieves a consistent performance gain over the baseline; 2) with fewer yet informative frames, we maintain gains over baseline while requiring only half the number of tokens and reducing the LLM time to 50% of the original one; 3) for keyframe extraction time, our approach saves 90% of the time compared to CoS (Hu et al. 2025a), yet achieves performance gain. This demonstrates our efficiency in handling long videos.

Qualitative Results. Fig. 5 demonstrates the visualization results of keyframe selection and model responses. Our trained temporal agent is shown to achieve two capabilities: basic object recognition, *e.g.*, "short-haired woman" or "airplane", and temporal event relationship comprehension, *e.g.*, "entering the museum" or "appeared before". When the short-haired woman is localized, preceding contextual frames (the first scenes that the woman visits) are simultaneously captured, which confirms that our TSPO effectively guides the temporal agent to learn complex query-event correlation capacity. Furthermore, our precise keyframe selection enables the MLLM to generate accurate responses.

Conclusion

This paper proposes a Temporal Sampling Policy Optimization framework, which addresses the unsupervised and non-differentiable challenge of sparse frame sampling in Video-MLLMs. We propose an RL framework to optimize sparse frame sampling in an end-to-end manner, and propose a TSPO-targeted training data construction pipeline. Extensive comparison experiments and ablation studies validate the effectiveness and generalizability of our method.

Acknowledgments

This work was supported in part by National Science and Technology Major Project under Grant 2023ZD0121300, National Natural Science Foundation of China under Grants 62088102, U24A20325, 12326608, 62522102 and 62373043, and Key Research and Development Plan of Shaanxi Province under Grant 2024PT-ZCK-80.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cheng, C.; Xu, T.; Feng, Z.; Wu, X.; Li, H.; Zhang, Z.; Atito, S.; Awais, M.; Kittler, J.; et al. 2025. One Model for ALL: Low-Level Task Interaction Is a Key to Task-Agnostic Image Fusion. *arXiv preprint arXiv:2502.19854*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Cui, W.; Du, S.; Yao, R.; Tang, C.; Ye, A.; Wen, F.; and Tian, Z. 2023. RDD: Learning Reinforced 3D Detectors and Descriptors Based on Policy Gradient. *IEEE Transactions on Multimedia*.

DeepSeek-AI, D. G.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Guo, W.; Chen, Z.; Wang, S.; He, J.; Xu, Y.; Ye, J.; Sun, Y.; and Xiong, H. 2025. Logic-in-Frames: Dynamic Keyframe Search via Visual Semantic-Logical Verification for Long Video Understanding. *arXiv preprint arXiv:2503.13139*.

Hu, J.; Cheng, Z.; Si, C.; Li, W.; and Gong, S. 2025a. CoS: Chain-of-Shot Prompting for Long Video Understanding. *arXiv preprint arXiv:2502.06428*.

Hu, K.; Gao, F.; Nie, X.; Zhou, P.; Tran, S.; Neiman, T.; Wang, L.; Shah, M.; Hamid, R.; Yin, B.; et al. 2025b. M-LLM Based Video Frame Selection for Efficient Video Understanding. *arXiv preprint arXiv:2502.19680*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kim, J.; Kim, H.; Lee, H.; and Ro, Y. M. 2024. SALOVA: Segment-Augmented Long Video Assistant for Targeted Retrieval and Routing in Long-Form Video Analysis. *arXiv preprint arXiv:2411.16173*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, R.; Wang, X.; Zhang, Y.; Wang, Z.; and Yeung-Levy, S. 2025. Temporal Preference Optimization for Long-Form Video Understanding. *arXiv preprint arXiv:2501.13919*.

Liang, J.; Meng, X.; Wang, Y.; Liu, C.; Liu, Q.; and Zhao, D. 2024. End-to-end video question answering with frame scoring mechanisms and adaptive sampling. *arXiv preprint arXiv:2407.15047*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, H.; Feng, J.; Kong, L.; Liang, N.; Tao, D.; Chen, Y.; and Zhang, M. 2023. One for all: Towards training one graph model for all classification tasks. *arXiv preprint arXiv:2310.00149*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, R.; Li, C.; Tang, H.; Ge, Y.; Shan, Y.; and Li, G. 2024b. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, 1–18. Springer.

Liu, R.; Tang, H.; Liu, H.; Ge, Y.; Shan, Y.; Li, C.; and Yang, J. 2024c. Ppllava: Varied video sequence understanding with prompt guidance. *arXiv preprint arXiv:2411.02327*.

Liu, Z.; Dong, Y.; Liu, Z.; Hu, W.; Lu, J.; and Rao, Y. 2025. Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution. arXiv:2409.12961.

Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; et al. 2024d. NVILA: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

OpenAI. 2023. GPT-4V. https://openai.com/index/gpt-4v-system-card/.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pu, Y.; Wu, X.; Yang, L.; and Wang, S. 2024. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Schulman, J. 2020. Approximating KL Divergence.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, X.; Xiong, Y.; Zhao, C.; Wu, L.; Chen, J.; Zhu, C.; Liu, Z.; Xiao, F.; Varadarajan, B.; Bordes, F.; Liu, Z.; Xu, H.; J. Kim, H.; Soran, B.; Krishnamoorthi, R.; Elhoseiny, M.; and Chandra, V. 2024. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. *arXiv preprint arXiv:2410.17434*.

Tang, X.; Qiu, J.; Xie, L.; Tian, Y.; Jiao, J.; and Ye, Q. 2025. Adaptive Keyframe Sampling for Long Video Understanding. *arXiv preprint arXiv:2502.21271*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; et al. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Gu, X.; Huang, S.; Xu, B.; Dong, Y.; et al. 2024c. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

Wang, Y.; Zhang, R.; Wang, H.; Bhattacharya, U.; Fu, Y.; and Wu, G. 2023. Vaquita: Enhancing alignment in llm-assisted video understanding. *arXiv preprint arXiv:2312.02310*.

Wei, T.; Patel, Y.; Shekhovtsov, A.; Matas, J.; and Barath, D. 2023. Generalized differentiable RANSAC. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17649–17660.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*.

Xu, J.; Rao, Y.; Zhou, J.; and Lu, J. 2024a. Procedure-aware action quality assessment: Datasets and performance evaluation. *International Journal of Computer Vision*, 132(12): 6069–6090.

Xu, J.; Rao, Y.; Zhou, J.; and Lu, J. 2025a. Transferable Unintentional Action Localization With Language-Guided Intention Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5): 3863–3877.

Xu, J.; Yin, S.; and Peng, Y. 2025. Human-Centric Fine-Grained Action Quality Assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8): 6242–6255.

Xu, J.; Zhang, Y.; Zhou, W.; and Liu, H. 2025b. BFSTAL: Bidirectional Feature Splitting with Cross-Layer Fusion for Temporal Action Localization. *IEEE Transactions on Circuits and Systems for Video Technology*.

Xu, M.; Gao, M.; Gan, Z.; Chen, H.-Y.; Lai, Z.; Gang, H.; Kang, K.; and Dehghan, A. 2024b. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*.

Ye, J.; Wang, Z.; Sun, H.; Chandrasegaran, K.; Durante, Z.; Eyzaguirre, C.; Bisk, Y.; Niebles, J. C.; Adeli, E.; Fei-Fei, L.; et al. 2025. Re-thinking Temporal Search for Long-Form Video Understanding. *arXiv preprint arXiv:2504.02259*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Yu, S.; Jin, C.; Wang, H.; Chen, Z.; Jin, S.; Zuo, Z.; Xu, X.; Sun, Z.; Zhang, B.; Wu, J.; et al. 2024. Frame-Voyager: Learning to Query Frames for Video Large Language Models. *arXiv preprint arXiv:2410.03226*.

Zhang, K.; Li, B.; Zhang, P.; Pu, F.; Cahyono, J. A.; Hu, K.; Liu, S.; Zhang, Y.; Yang, J.; Li, C.; and Liu, Z. 2024a. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models. arXiv:2407.12772.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024b. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Zhang, R.; Gui, L.; Sun, Z.; Feng, Y.; Xu, K.; Zhang, Y.; Fu, D.; Li, C.; Hauptmann, A.; Bisk, Y.; et al. 2024c. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024d. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Xiao, S.; Yang, X.; Xiong, Y.; Zhang, B.; Huang, T.; and Liu, Z. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264*.

Zohar, O.; Wang, X.; Dubois, Y.; Mehta, N.; Xiao, T.; Hansen-Estruch, P.; Yu, L.; Wang, X.; Juefei-Xu, F.; Zhang, N.; Yeung-Levy, S.; and Xia, X. 2024. Apollo: An Exploration of Video Understanding in Large Multimodal Models. *arXiv preprint arXiv:2412.10360*.