Attend and Replay: Efficient Action Understanding in Long Videos via Mechanistic Interpretability

Puyue Hou[1,2] Jinjin Zhang[1,2] Di Huang[1,2*]

[1]State Key Laboratory of Complex and Critical Software Environment, Beihang University
[2]School of Computer Science and Engineering, Beihang University
Beijing 100191, China

{hhhhhpy06, jinjin.zhang, dhuang}@buaa.edu.cn

Abstract

Efficient action understanding in long videos remains a significant challenge for multimodal large language models (MLLMs), primarily due to the difficulty in localizing target actions within long frame sequences. This stems from the overwhelming interference of unrelated actions in long videos. In this work, we focus on efficient action localization to enhance video understanding from an internal interpretability perspective, leveraging the intricate relationship between text and video tokens to remove irrelevant tokens. By tracing attention distribution across videos of varying frame lengths, we observe that unsuccessful action understanding directly correlates with unrelated actions that receive notable attention scores. Motivated by these findings, we propose an Attend and Replay method that efficiently locates critical action information and strengthens its semantic representation. This approach first reduces unrelated action tokens using an attention-guided spatiotemporal pruning strategy, then enriches target action tokens via a pivot-token aggregation method. Extensive experiments show that integrating our method with existing MLLMs (e.g., LLava-Video, Qwen2.5-VL, MiMo-VL) achieves superior performance against other counterparts on various datasets, while enjoys lightning inference speed.

1. Introduction

Recent advances in multimodal large language models (MLLMs) have demonstrated strong performance in video understanding tasks [13, 29], yet they remain limited in long-video comprehension, particularly in interpreting actions within specific contexts [4, 8, 13–15, 29, 38]. This challenge arises from the extreme imbalance between target actions and lengthy video content, where hundreds of sequential actions complicate precise localization. For example, answering "What is the student doing while the athlete is competing in the indoor stadium?" in a one-hour vlog requires identifying relevant spatiotemporal segments and inferring contextualized actions, demanding both fine-grained motion discrimination and query-guided attention. Current limitations of MLLMs reveal a critical bottleneck in adaptive long-context video processing, underscoring action understanding as fundamental for advancing long-video comprehension.

In recent work, several studies have investigated attention patterns in MLLMs on image tasks [6, 17], revealing that whether an MLLM concentrates its attention on visual regions tightly related to the question strongly affects performance. By contrast, the interpretation mechanisms for video inputs, especially for the challenging problem of action understanding in long videos remains largely unexplored. Through qualitative analysis and systematic ablations that independently vary the number of key and irrelevant frames, we show that the completeness of key-frame information is crucial for correct answers, while an increasing fraction of irrelevant frames consistently degrades accuracy. Additionally, we analyze the internal attention patterns of MLLMs to probe the interaction between video and text. Our analysis considers three aspects: (1) attention allocation between text and video tokens, (2) the relationship between key-frame attention and prediction accuracy, and (3) how attention to top-ranked frames correlates with outcomes. We observe a dominant allocation of attention to text tokens that tends to suppress visual signals. Moreover, although correct predictions receive only limited absolute attention, their attention is more narrowly concentrated on key frames; incorrect predictions, by contrast, exhibit a more uniform attention spread across frames.

Guided by these findings, we propose Attend and Re-

*Corresponding author

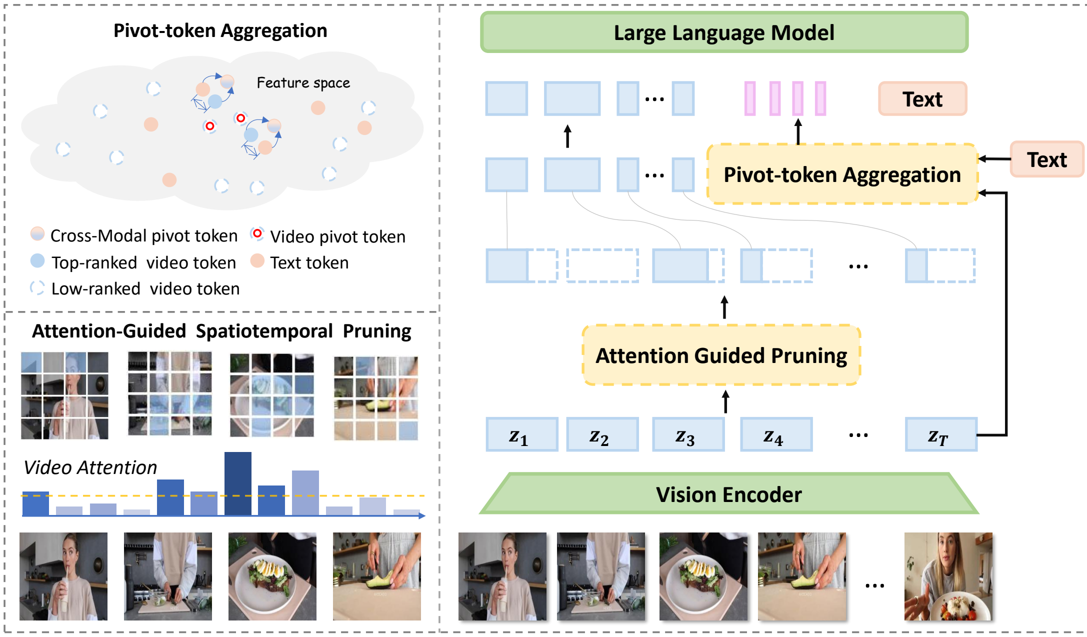


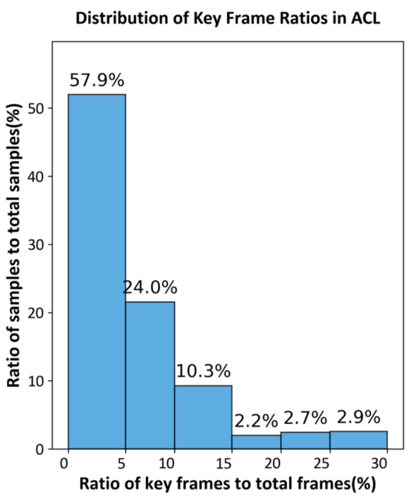
Figure 1. The Attend and Replay Framework. This training-free method comprises two critical components: attention-guided spatiotemporal pruning and pivot-token aggregation. The former reduces the number of video tokens by pruning those with low attention scores, which were found to contain the majority of irrelevant information. The latter employs additional pivot tokens to enhance target action semantics using a minimal number of tokens: cross-modal pivot tokens inject critical textual cues by interpolating top-ranked video tokens with their most semantically similar text tokens, while video pivot tokens supplement action information by adding extra visual tokens drawn from top-ranked frames when the attention distribution becomes overly uniform.

play, a training-free method that efficiently locates critical action information and strengthens its semantic representation prior to feeding tokens into the LLM component. Attend and Replay first applies an attention-guided spatiotemporal pruning strategy to discard irrelevant frame tokens, in line with our observation that such tokens attract little attention. It then enriches target action tokens via a pivot-token aggregation scheme. We introduce two pivot types: (i) additional video tokens drawn from top-ranked frames to increase attention concentration when the distribution is overly uniform, and (ii) text-like pivot tokens retrieved as the most semantically similar text tokens to top-ranked video tokens and interpolated with visual tokens to inject critical semantic cues. We evaluate our method with LLaVA-Video [38], Qwen2.5-VL [2], and MIMO-VL [22] on a dedicated Action Long (ACL) dataset and two long-video benchmarks, VideoMME and LongVideoBench [10, 26]. Our approach matches or surpasses the original models (without token reduction) on ACL while cutting inference time to roughly 40% of the baseline. On the two long-video benchmarks, Attend and Replay achieves higher accuracy and faster inference than state-of-the-art training-free pruning methods, demonstrating strong generalization and robustness.

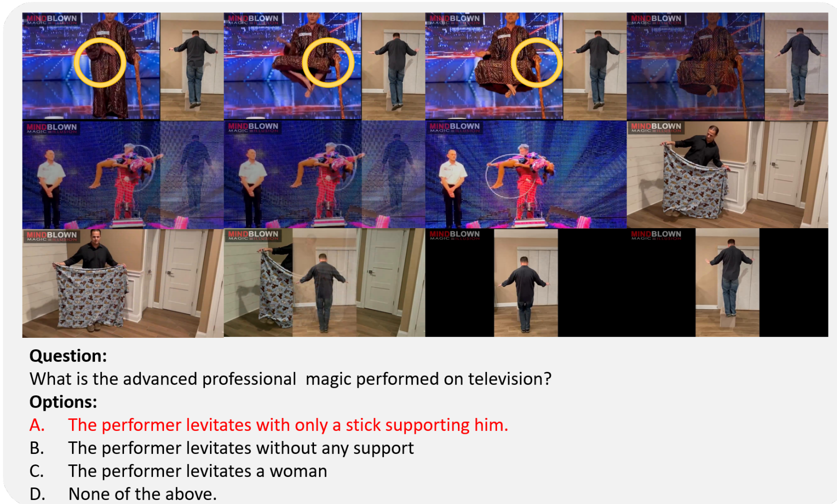
2. Related Works

Long-video Understanding. Long-video understanding has emerged as an important research direction [40]. Unlike standard video-understanding tasks, long-video problems often involve hour-long content and thus pose greater challenges for MLLMs. Representative benchmarks cover both narrative-level and localized recognition or reasoning tasks [3, 9, 10, 23, 26, 39]. To tackle these challenges, some works improve core capabilities via large-scale pretraining and fine-tuning [2, 7, 8, 18, 22, 27, 31, 37], while others adopt token-compression paradigms to reduce computational cost, either during training or in a training-free manner [11, 12, 19–21, 24, 30, 32, 33]. In this paper, we concentrate on action understanding in long videos and introduce a training-free method to strengthen long-video comprehension.

Mechanistic Interpretability. Understanding the internal mechanisms of MLLMs is a central research goal. Early work on mechanistic interpretability examined attention patterns in LMs [1, 16, 28, 34, 35]. More recent studies probe MLLMs directly: [6] analyze attention in spatial reasoning, [17] trace the evolution of visual-token representations across layers and the process of visual–text integration, and [36] show that many visual tokens exert their



(a) Distribution of key-frame ratios in ACL.



(b) Example long-video sample with key frames highlighted.

Figure 2. **Left:** Distribution of key-frame ratios in ACL, showing that over half of samples have key frames occupying $< 5\%$ of the video. **Right:** An example long-video question; highlighted frames are key frames relevant to the question.

primary influence in early LLM layers. Our work extends these analyses to the understudied setting of action understanding in long videos, investigating how attention and token interactions govern performance.

3. Key Frame Analysis in Long Videos

In this section, we systematically analyze the impact of key frames and irrelevant frames on question answering in long video scenarios. Our goal is to address the following two questions: ***How essential are key frames for answer correctness? What influence do irrelevant frames introduce to the answer?*** To this end, We first construct a dataset consisting of approximately 600 question–answer pairs derived from long videos, each paired with precise key frame annotations. Subsequently, we conduct a series of experiments to investigate the influence of key frames, aiming to gain deeper insights into action understanding in long videos.

3.1. ACL Benchmark

To investigate impact of key frames and irrelevant frames, we construct a benchmark specifically designed for action understanding in long videos with key frame annotations. We select two widely-used long video benchmarks, VideoMME [10] and LongVideoBench [26], as our source datasets. We filter out the samples for action understanding and propose a coarse-to-fine strategy to progressively locate the key frames with the aid of a Large Language Model (LLM). We adopt the original format ⟨video, question⟩ from the source datasets and repurpose the typical action understanding samples from VideoMME and LongVideoBench to define our action understanding task. Additionally, to ensure compliance with the definition of long videos, we filter out all video samples shorter than five minutes. Since the two benchmarks lack key frame annotations, we propose an automatic pipeline to generate key frames. A key frame is expected to satisfy the following three criteria: (1) high relevance to the question, (2) completeness of the involved actions and scenes, and (3) uniqueness in contrast to irrelevant video content. Based on these principles, we design a coarse-to-fine strategy to progressively localize the key frames with the aid of a LLM.

We densely sample video frames to avoid information loss, then split them into equal-length segments as candidate regions. Each segment is processed by an MLLM with the question to generate an answer. Since only the true key frame should yield the correct answer, we retain only cases where exactly one segment succeeds, narrowing down the key frame's temporal range. If no segment produces the correct answer, we refine segmentation and repeat the process hierarchically. To precisely locate the key frame within the coarse range, we use LLMs (e.g., DeepSeek) for reasoning. Each frame in the segment is captioned by an MLLM, and the captions, question, and ground-truth answer are fed into an LLM to determine the best-matching key frame based on three criteria. Finally, human verification ensures localization accuracy. Through this pipeline, we construct a long-video dataset with precise key frame annotations, focused on action understanding. We name this dataset ActionLong, and refer to it as **ACL** throughout the paper.

In the following experiment, we use Qwen2.5-VL-7B [2] as the evaluation model and follows [10, 26] to utilize the selection accuracy as the metric. In Fig. 2, we display the key frame ratio distribution and a specific action under-

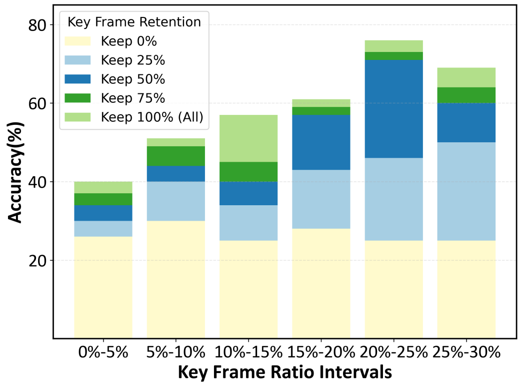


Figure 3. Impact of key frame removal on ACL performance. We group the samples by their key frame length and evaluate the performance changes shown in each bar. We keep all the input frames except removing key frames by the experiment settings.

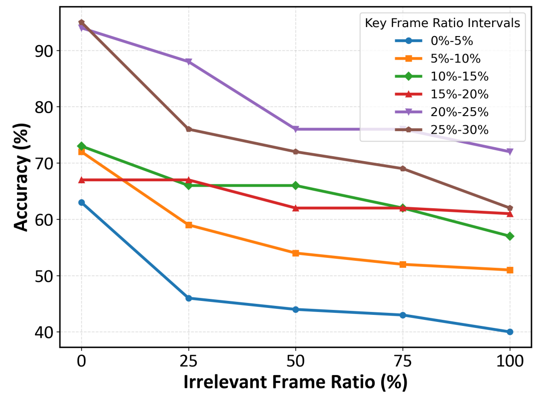


Figure 4. Impact of Irrelevant Frames on ACL Performance. To assess the impact of irrelevant frames, we preserve all key frames and incrementally add irrelevant frames at varying ratios. The irrelevant frames are added along the temporal dimension. For example, adding 25% irrelevant frames means appending the first 25% of the whole irrelevant frames, while adding 100% denotes including all irrelevant frames sampled from the original video.

standing samples. The sample show that the key frames contains the critical information for answering the question and other irrelevant frames might intervene MLLMs to locate and understand the key frame information. For key frames, it is clearly shown that over a half samples' key frames only occupy less than 5% proportion of the whole video and all samples' key frames occupy less than 30% proportion of the whole video. Such scarce distribution indicates difficulty of localizing the key frames in long videos.

3.2. How essential are key frames to answer correctness?

Although the crucial role of key frames in model evaluation is widely acknowledged, quantitatively assessing their importance has been largely overlooked in previous studies. To thoroughly investigate their contribution, we conduct an ablation study by varying the number of key frames in the input and evaluating the resulting impact on prediction accuracy. As shown in Fig. 3, we observe a consistent trend: as more key frames are removed (i.e., lower retention percentages), the accuracy drops significantly—especially in groups with a higher proportion of key frames. This indicates that key frames contain substantial information critical for understanding and answering video-related questions. When all key frames are removed, the accuracy across all groups drops to approximately 25%. We conjecture that this phenomenon results from random guessing by the model, as the remaining input frames no longer provide useful information.

Furthermore, we find that merely retaining all key frames does not necessarily yield high performance—particularly when the proportion of key frames is low. For example, when the key frame ratio falls below 5%, the prediction accuracy drops below 50%. This raises an important question: Is the inclusion of key frames sufficient to guarantee answer correctness? We explore this question in more detail through targeted experiments in the next section.

3.3. What influence do irrelevant frames introduce to the answer?

As the previous experiment demonstrates that the inclusion of all key frames does not lead to the expected performance, we hypothesize that this may be caused by interference from irrelevant frames. To verify this hypothesis, we conduct an ablation study by progressively adding irrelevant frames and evaluating the model's performance on the ACL. As presented in Fig. 4, it clearly reveals a consistent inverse relationship: as the proportion of irrelevant frames increases, accuracy consistently declines. This indicates that the presence of irrelevant content has a detrimental effect on model performance.

Although all key frame ratio groups exhibit a downward trend in performance, their initial accuracy levels vary. For example, key frame density intervals such as 20%–25% and 25%–30% maintain relatively higher accuracy, whereas the 0%–5% group experiences a much sharper decline. This indicates that merely exposing the model to key frames does not guarantee 100% accuracy. We conjecture that this phenomenon arises from two main factors. First, due to the extremely long temporal length of the videos, sampling only a few hundred frames inevitably leads to the loss of critical action-related information. Second, the limited action understanding capabilities of current MLLMs may also contribute to incorrect predictions. These findings highlight the critical importance of minimizing irrelevant information in the input to ensure higher model accuracy.

4. Attention Distribution Analysis in Long Videos

In this section, we take a further step to explore how MLLMs locates and interprets and answer the question from the lens of attention mechanism. We hypothesize that an MLLM's ability to allocate majority attention to key frame regions is a critical factor in deriving correct answers. To test this hypothesis, we analyze attention distribution through three lenses: (1) the mechanism of attention allocation between text and video tokens, (2) the correlation between key frame attention distribution and prediction accuracy, and (3) the relationship between top-ranked frame attention and prediction confidence.

4.1. Text vs. Video Attention Analysis

We first define the attention score of a given video token. Given a video input $\mathcal{V} \in \mathbb{R}^{T \times H \times W \times C}$, a visual encoder produces $T \times N$ video tokens $\mathbf{z}^{(0)} \in \mathbb{R}^{T \times N \times d}$, which are concatenated with text tokens into a unified token sequence $\mathbf{X}^{(0)} \in \mathbb{R}^{(T \times N + M) \times d}$ that serves as input to a LLM with L layers. At each layer $\ell = 1, \ldots, L$, the model computes multi-head self-attention, which can be expressed as:

$$N_h^{(l,h)} = \mathcal{A}^{(l,h)}\mathbf{V} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}} + \mathbf{M}\right)\mathbf{V}, \quad (1)$$

where h indicates the h-th head, and N_h is the output of head, $\mathbf{Q} = \mathbf{Z}^{(\ell-1)}W^Q$, $\mathbf{K} = \mathbf{Z}^{(\ell-1)}W^K$, $\mathbf{V} = \mathbf{Z}^{(\ell-1)}W^V$. In this paper, we use post-softmax attention $A^{(l,h)}$ and average the score across all attention heads at layer l, which can be expressed as:

$$\mathcal{A}^{(l)} = \frac{1}{H}\sum_{h=1}^{H} \mathcal{A}^{(l,h)}. \quad (2)$$

In Fig. 5, we focus on the attention scores from the first generated token. It reveals a striking disparity between token length distribution and attention allocation. This pronounced attention bias toward text tokens demonstrates their critical role in action understanding. The results demonstrate a significant imbalance in attention allocation, with text tokens receiving approximately 90 times more attention than video tokens. This skewed attention pattern severely hinders the model's ability to interpret visual information, highlighting the urgent need to improve attention distribution toward video content.

4.2. Key Frame Attention Distribution Analysis

Based on Eq. 2, we select the attention score of the video tokens as $\mathcal{A}_v^{(l)} \in \mathbb{R}^{(TN) \times 1}$. To characterize the attention distribution of key frame tokens, we propose a histogram-based density estimation method. For each key frame token,

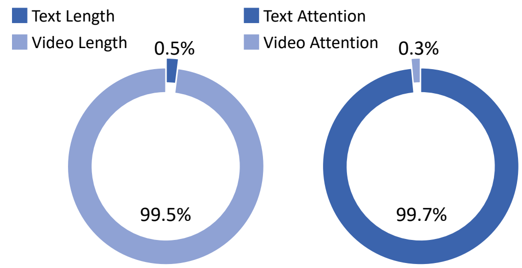


Figure 5. Text vs. Video Attention attention. The dominance of text tokens highlights their critical role in understanding action in long videos.

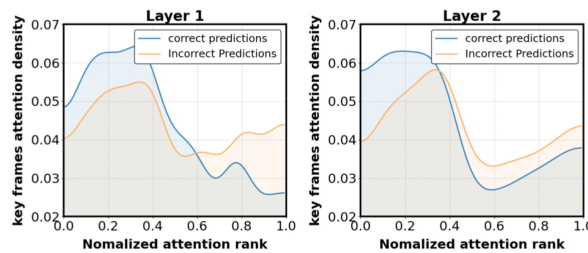


Figure 6. Key frame attention distribution of two groups: the samples with correct predictions and samples with incorrect predictions. We utilize discrete probability mass function (PMF) to characterize the attention distribution. The PMF values are smoothed using a Gaussian kernel with bandwidth $h = 5.0$ for visual clarity,. Correct predictions exhibit a markedly higher concentration of attention in top-ranked positions, whereas incorrect predictions display lower attention density at the top and a more uniformly dispersed attention distribution.

we collect its attention scores in the sequence and sort them in descending order.

Let $\tilde{\mathcal{A}}_k^{(l)} = [a_{(1)}^{(l)}, a_{(2)}^{(l)}, \ldots, a_{(M)}^{(l)}]$ denote the sorted attention scores, where $a_{(1)}^{(l)} \geq a_{(2)}^{(l)} \geq \cdots \geq a_{(M)}^{(l)}$ and their responding index $\mathbf{I}_k^l = [i_{(1)}^{(l)}, i_{(2)}^{(l)}, \ldots, i_{(M)}^{(l)}]$. M is the key frame token length. A normalized histogram is computed over these sorted scores using B fixed bins. The discrete probability mass function (PMF) can be expressed as:

$$\mathbf{h}^{(l)} = \left[h_1^{(l)}, h_2^{(l)}, \ldots, h_B^{(l)}\right], \quad (3)$$

where each bin probability is calculated as:

$$h_r^{(l)} = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}\left\{i_{(j)}^{(l)} \in [b_{r-1}, b_r)\right\}. \quad (4)$$

Here $\mathbb{1}$ is the indicator function, and $h_r^{(l)}$ represents the empirical probability of attention scores for tokens falling in bin r. The aggregate attention distribution $\mathbf{H}$ is obtained by averaging PMFs over N samples:

$$\mathbf{H} = [H_1, H_2, \ldots, H_B], \tag{5}$$

where:

$$H_r = \frac{1}{N}\sum_{i=1}^{N} h_r^{(l)}. \tag{6}$$

For visual clarity, the discrete PMF is represented using a continuous smoothing function and the results are shown in Fig. 6. It shows correct predictions exhibit a pronounced peak at top-ranked positions (left side), indicating that attention is highly concentrated on a small set of high-importance key frames, whereas incorrect predictions display a lower peak and a flatter, more uniform distribution with relatively more mass in later ranks, suggesting attention is dispersed across less relevant frames. This pattern implies that concentrating attention on truly informative key frames correlates with higher accuracy. Here We only display the PMF of the first two layers for brevity. We find the same property in other layers.

4.3. Top-ranked Frame Attention Distribution Analysis

As we find a more concentrated attention distribution of correct predicted samples and more dispersed attention distribution of incorrect predicted samples on key frames, we further explore whether the property exists in the complete frames. Specifically, we quantify frame-wise attention distribution in the top-ranked tokens. For all video tokens with attention scores $\mathcal{A}_v^{(l)} \in \mathbb{R}^{(TN)\times 1}$, we select the top-$k$ attended tokens ($k = 30\%$), and their corresponding index $\mathbf{D}_k^l = [d_{(1)}^{(l)}, d_{(2)}^{(l)}, \ldots, d_Q^{(l)}]$. We statistics the number of tokens belonging to each frame, which can be expressed as:

$$\mathbf{S}^{(l)} = \left[t_1^{(l)}, t_2^{(l)}, \ldots, t_T^{(l)}\right], \tag{7}$$

where

$$t_p^{(l)} = \sum \mathbb{1}\left\{d_j^l \in \mathcal{T}_p\right\}, \tag{8}$$

$t_p^{(l)}$ represents the number of tokens belonging to p-th frame. Finally, we normalize and calculate standard deviation across $\mathbf{S}^{(l)}$:

$$c^l = \mathbf{sd}(\mathbf{S}^{(l)}). \tag{9}$$

Finally, we statistics standard deviation distribution of all samples by using histogram which is expressed in Eq. 5 and Eq. 6 and the visualization result are shown in Fig. 7. It reveals a clear dichotomy in attention distribution patterns between correct and incorrect predictions: correct predictions exhibit substantially higher standard deviations in their top-ranked frame distributions, indicating highly concentrated attention on critical frames; conversely, incorrect predictions show markedly lower dispersion, reflecting uniform attention allocation across frames. This divergence demonstrates that successful predictions correlate strongly with focused attention mechanisms, whereas prediction failures align with undiscriminating attention spread, confirming that attention concentration serves as a key indicator of model decision confidence.

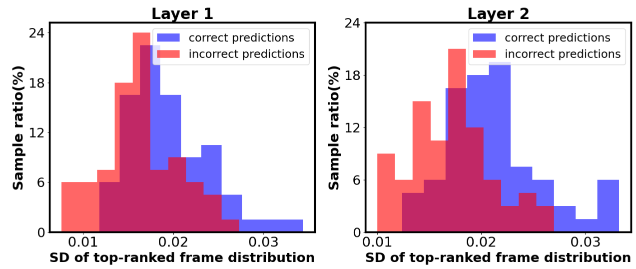


Figure 7. Dispersion of top-ranked frame distribution in layer 1 of two groups: the samples with correct predictions and samples with incorrect predictions. We here utilize the stand deviation of top-ranked frame distribution to characterize the attention dispersion. Higher standard deviation represents a more concentrated attention on specific frames while lower standard deviation reflects uniform attention allocation across frames. Correct predictions exhibit higher standard deviations and incorrect predictions show markedly lower dispersion, demonstrating the different attention pattern relates to different prediction results.

5. Method

Based on the experiments and analysis in Section 3 and Section 4, we observe two distinct properties of action understanding in long videos: (1) the model prediction is easy to be influenced by information from irrelevant frames. (2) Correct predictions and incorrect predictions exhibit different attention distribution pattern that the former tends to apply a concentrated attention allocation on key frames while the latter prefers uniformly focusing on complete frames whether it is key frame or not. Motivated by the observations above, we propose an **Attend and Replay** method that aims to remove the irrelevant frames and enrich the key frame information to enhance the action understanding performance in long videos while enjoys a lighting speed. The pipeline is shown in Fig. 1. Specifically, we first propose a text-guided pruning strategy to reduce irrelevant video tokens, then enriches target action tokens via a pivot-token aggregation method.

5.1. Attention-guided Spatiotemporal Pruning

We introduce an Attention-guided Spatiotemporal token pruning method to reduce irrelevant video tokens. As shown in Fig. 6, we observe the key frame token is less distributed in the low attention score ranges in the LLM layers, we apply an attention-guided token pruning method that pruning the video tokens that exhibit low attention scores.

We concatenate the video tokens and text tokens into $\mathbf{W} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T, \mathbf{y}] \in \mathbb{R}^{(O+M)\times D}$ where $O = \sum_{t=1}^{T} K_t$, and M is the length of text tokens. After processing through the first transformer layer of LLM, we obtain attention scores to the final token by Eq. 2 as $\mathcal{A}^{(1)}$: We retain only the top-k tokens:

$$\mathcal{I} = \underset{i\in[1,O]}{\text{top-k}}\,(\mathcal{A}^{l}), \tag{10}$$

$$\mathbf{W}_{\text{pruned}} = \mathbf{W}[\mathcal{I}] \in \mathbb{R}^{k\times d}. \tag{11}$$

The pruned sequence $\mathbf{W}_{\text{pruned}}$ propagates through subsequent layers for final prediction.

5.2. Pivot-token Aggregation

In Section 3 and Section 4, we observe text tokens accounts for the majority of attention, indicating the their importance. Meanwhile, the attention distribution of correct prediction samples differs from that of incorrect prediction samples, which is a distinguishable signal for detecting the incorrect predictions. Motivated by the observations, we propose a pivot-token aggregation method to enrich the necessary video information. We include two kinds of pivot tokens: Video Pivot Tokens and Cross-Modal Pivot tokens.

Video Pivot Tokens. For pruned tokens $\mathbf{W}_{\text{pruned}} \in \mathbb{R}^{k\times d}$, let $\mathbf{D}^{l}$ denotes their corresponding index $\mathbf{D}_k = [d_1, d_2, \dots, d_k]$. We compute frame attention statistics:

$$\alpha_t = \sum \mathbb{1}\left\{d_i \in \mathcal{T}_t\right\}. \tag{12}$$

When the standard deviation σ obtained by Eq. 9 less than a threshold γ, we consider it as a uniform distribution that is likely to lose important information, thus we augment it by adding additional tokens from top-K frames:

$$\mathcal{T}_{\text{top}} = \underset{t\in[1,T]}{\text{top-}K}(\alpha_t), \tag{13}$$

$$\delta_t^{\text{vis}} = \underset{d^{l}\in\mathcal{T}_t n \mathbf{W}_{pruned}}{\text{top-}\lceil r\cdot|\tilde{\mathbf{z}}t|\rceil}\,(\mathcal{A}^{(l)}(\mathbf{d})) \quad \forall t \in \mathcal{T}_{\text{top}}\,, \tag{14}$$

$$\mathcal{P}_t^{\text{vis}} = \mathbf{W}[\delta_t^{\text{vis}}]\,, \tag{15}$$

$$\mathbf{P}^{\text{vis}} = [\mathcal{P}_1^{\text{vis}}, \mathcal{P}_2^{\text{vis}}, ..., \mathcal{P}_K^{\text{vis}}], \tag{16}$$

where r controls the retention ratio of each frame.

Cross-Modal Pivot Tokens. For video token $\mathbf{v}_i \in \mathbf{W}_{\text{pruned}}$ with top-β attention scores, and text tokens $\mathbf{y} \in \mathbb{R}^{M\times d}$, we compute the cosine similarity between each video token and text token:

$$\mathbf{\Gamma}_{ij} = \frac{\mathbf{v}_i\mathbf{y}_j^{\top}}{|\mathbf{v}_i||\mathbf{y}_j|}, \tag{17}$$

we then choose the text token with the highest similarity score for each video token as $\mathbf{y}_{ij^*}$, we consider those text tokens carry discriminative information, and we fuse them as the pivot tokens:

$$\mathcal{P}_i^{\text{text}} = \lambda\mathbf{v}_i + (1-\lambda)\mathbf{y}_{ij^*}, \tag{18}$$

$$\mathbf{P}^{\text{text}} = [\mathcal{P}_1^{\text{text}}, \mathcal{P}_2^{\text{text}}, ..., \mathcal{P}_\beta^{\text{text}}]. \tag{19}$$

The final representation combines pruned tokens with both pivot types:

$$\mathbf{W}_{\text{final}} = \mathbf{W}_{\text{pruned}} \cup \mathbf{P}^{\text{vis}} \cup \mathbf{P}^{\text{text}}. \tag{20}$$

This augmented sequence propagates through subsequent transformer layers.

6. Experiments

6.1. Experimental Settings

Benchmarks. We evaluate our method on three long video benchmarks: ACL, VideoMME [10] and LongVideoBench [26]. ACL is specifically designed for action understanding in long videos, while VideoMME and LongVideoBench are comprehensive video evaluation benchmark. We apply our method for the above three benchmark to validate the generaliztion ability of our method. They all utilize multiple-choice questions for evaluation and we utilize prediction accuracy as the metric.

Models. We evaluate our approach across three representative MLLMs from diverse model families: LLaVA-Video-7B [38], Qwen2.5VL-7B [2] and MIMO-VL-7B [22]. LLaVA-Video-7B is specifically designed for video-related tasks while Qwen2.5VL-7B and MIMO-VL-7B are universal MLLMs for all visual-related tasks.

Baselines. We compare our method with state-of-the-art token compression baselines, FastV [5] ans DART [25]. We reproduce the results by using their official implementations.

Implementation details. We utilize 2 NVIDIA L20 GPUs for all experiments. we set the threshold γ for adding video pivot token as 0.2, the upper limit β for video pivot token numbers as 0.01, the fusion coefficient λ for cross-model pivot token as 0.8.

6.2. Main Results

We compare our method with other state-of-the-art token pruning method in Table 1. It reveals the effectiveness of our method, with particularly strong results on ACL, a benchmark specifically designed for action understanding in long videos, which is the primary focus of our method. Notably, for both Qwen2.5-VL-7B and Mimo-VL-7B, our approach not only reduces latency significantly but also yields accuracy gains, demonstrating that our method does not simply trade accuracy for efficiency but instead enhances both. This improvement stems from our analysis of the attention distribution in ACL, which guides targeted pruning and the aggregation of pivot tokens, enabling

Model	Frame	Method	ACL		VideoMME		LongVideoBench	
			Acc(%)(↑)	Latency(s)(↓)	Acc(%)(↑)	Latency(s)(↓)	Acc(%)(↑)	Latency(s)(↓)
Llava-Video-7B	128	Original	52.5	8.7	64.8	8.7	52.4	9.3
	128	FastV	36.6	5.1	52.1	5.4	45.8	5.8
	128	DART	48.9	4.6	**62.6**	4.8	50.3	5.3
	128	Ours	**52.0**	4.4	58.5	4.4	**51.0**	4.7
Qwen2.5-VL-7B	128	Original	47.3	15.9	64.1	16.5	60.2	17.3
	128	FastV	47.1	11.1	63.9	13.0	58.6	13.6
	128	DART	41.4	10.4	61.5	11.5	56.5	11.7
	128	Ours	**49.1**	7.4	**64.3**	8.1	**59.0**	8.5
Mimo-VL-7B	128	Original	54.1	17.1	65.9	18.6	60.3	19.1
	128	FastV	53.4	11.1	64.7	13.3	**60.8**	13.6
	128	DART	49.1	10.4	62.6	11.6	56.9	11.9
	128	Ours	**55.8**	7.4	**65.3**	8.0	60.0	8.3

Table 1. Performance comparison across models and compression methods on three benchmarks. Latency time measures the prefilling duration per sample, as all tasks complete after prefilling. Original denotes baseline results without token pruning.

substantial latency reduction while reinforcing the contribution of informative frames. Although VideoMME and LongVideoBench are not directly aligned with ACL's task characteristics, our method consistently outperforms alternative approaches on these benchmarks as well, further validating its robustness and general effectiveness. These results collectively confirm that our method achieves a favorable speed-accuracy trade-off and offers a clear advantage over existing methods.

Method	Accuracy(s)	Latency(s)	Retention rate(%)
Original	47.3	15.9	100.0
+ASP	48.1	6.9	30.0
+PTA	48.9	7.4	30.1

Table 2. Ablation study on different components. ASP stands for the attention-guided spatiotemporal pruning and PTA stands for the pivot-token aggregation. We compute and average the latency on samples which need both types of pivot tokens that evaluate the upper bound of the latency time.

6.3. Ablation Study

We ablate the effectiveness of the attention-guided spatiotemporal pruning and pivot-token aggregation method, including both the accuracy and efficient evaluation. The results are shown in Table 2. Table 2 reports the ablation study on the two key components of our method, attention-guided spatiotemporal pruning (ASP) and pivot-token aggregation (PTA), showing their individual contributions to accuracy and efficiency. Compared with the original baseline, ASP alone significantly reduces latency from 15.9s to 6.9s while improving accuracy from 47.3% to 48.1%, demonstrating the effectiveness of pruning uninformative tokens guided by attention distribution while preserving salient ones. PTA further complements this by slightly increasing accuracy to 48.9% while keeping latency at a low 7.4s with a similar retention rate, indicating that aggregating pivot tokens strengthens the representation of critical information even under aggressive pruning. These results validate that both ASP and PTA are essential for the overall effectiveness of our approach, where ASP provides large efficiency gains and PTA reinforces accuracy, together enabling a superior balance of speed and performance.

7. Conclusion

In this paper, we addresses the critical challenge of long-video comprehension by revealing how the imbalance between key and irrelevant frames hinders action understanding in multimodal large language models. Through systematic analysis, we demonstrate that precise focus on key frames is essential for accurate predictions, while diluted attention leads to performance degradation. Building on these insights, we introduce Attend and Replay, a training-free token-level method that prunes irrelevant tokens and reinforces critical action semantics via pivot-token aggregation. Extensive experiments on multiple long-video benchmarks and action understanding tasks confirm that Attend and Replay not only achieves higher accuracy but also substantially reduces inference cost, offering an efficient and generalizable solution for advancing adaptive video comprehension.

8. Acknowledgement

This work is partly supported by the National Natural Science Foundation of China (82441024), the Research Program of State Key Laboratory of Critical Software Environment, and the Fundamental Research Funds for the Central Universities.

References

[1] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188, 2024. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 7

[3] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024. 2

[4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024. 1

[5] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 7

[6] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*, 2025. 1, 2

[7] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 2

[8] 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*, pages 24185–24198, 2024. 1, 2

[9] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. 2

[10] Chaoyou Fu, Yuhan Dai, Yongdong 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2, 3, 7

[11] Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large visual language models. *arXiv preprint arXiv:2501.01986*, 2024. 2

[12] Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. *arXiv preprint arXiv:2412.16117*, 2024. 2

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

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

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

[16] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023. 2

[17] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024. 1, 2

[18] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024. 2

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

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

[21] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18992–19001, 2025. 2

[22] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma,

Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025. 2, 7

[23] Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. *arXiv preprint arXiv:2503.12559*, 2025. 2

[24] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025. 2

[25] Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025. 7

[26] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 2, 3, 7

[27] Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation. *Advances in Neural Information Processing Systems*, 37:109922–109947, 2024. 2

[28] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 2

[29] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 1

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

[31] Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. *arXiv preprint arXiv:2504.17343*, 2025. 2

[32] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Rethinking temporal search for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8579–8591, 2025. 2

[33] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024. 2

[34] Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024. 2

[35] Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for llms. *arXiv preprint arXiv:2311.02262*, 2023. 2

[36] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*, 2025. 2

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

[38] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 2, 7

[39] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701, 2025. 2

[40] Heqing Zou, Tianze Luo, Guiyang Xie, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, Huaijian Zhang, et al. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*, 2024. 2