LENS: Adaptive Spatio-Temporal Zooming for Keyframe Sampling in Long-Form Videos

Ce Zhang Jinxi He Katia Sycara Yaqi Xie

Robotics Institute, Carnegie Mellon University
{cezhang, ginh, katia, yaqix}@cs.cmu.edu

Abstract. Despite rapid progress in Multi-modal Large Language Models (MLLMs), understanding long-form videos is still bottlenecked by limited context windows. While recent keyframe sampling methods attempt to mitigate this by distilling video inputs into a compact set of query-relevant frames, navigating the vast spatio-temporal search space remains challenging, as spatial detail and temporal coverage often conflict. To address this, we introduce LENS, a training-free keyframe sampling framework that dynamically decides when to zoom in for fine-grained details and when to zoom out for broader context based on the text query. Concretely, LENS adaptively allocates a limited frame budget between spatial zoom-ins, which highlight query-relevant regions within individual frames, and temporal zoom-outs, which expand the temporal scope through multi-frame aggregation, enabling the model to reason across multiple granularities while capturing both high-fidelity details and long-range context. Across diverse long-form video benchmarks, LENS consistently outperforms prior state-of-the-art keyframe sampling methods and delivers substantial gains over uniform sampling, improving Video-MME accuracy from 53.3% to 60.7% with Qwen2.5-VL. Code is available at https://github.com/zhangce01/LENS.

Keywords: Long-form Video Understanding · Keyframe Sampling · Multi-modal Large Language Models · Spatio-temporal Reasoning

1 Introduction

The emergence of Multi-modal Large Language Models (MLLMs) has expanded the scope of visual reasoning by bridging high-level linguistic understanding with complex visual perception [1, 19, 47, 50, 51, 53, 57]. By integrating robust vision encoders with the reasoning capabilities of large language models, these systems demonstrate strong generalization across a wide range of tasks, from zero-shot captioning to multi-modal dialogue. Despite these advances, extending MLLMs to long-form video understanding remains a significant challenge [14, 17, 20, 48, 49, 54]. Unlike images or short clips, long videos consist of thousands of frames with complex temporal dependencies, while current MLLMs can process only a limited number of visual tokens due to the limited context window. For example, a 30-minute video at 30 fps contains roughly 54,000 frames, which can translate

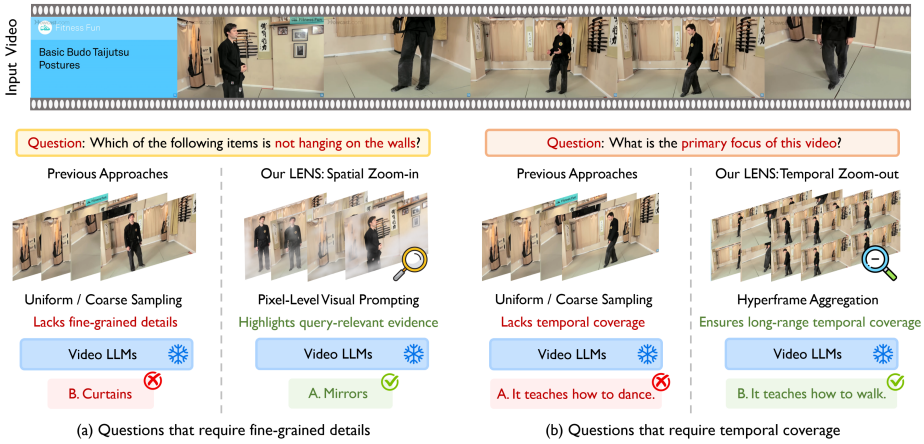

Fig. 1: Illustration of LENS for multi-granular video understanding. For detail-oriented questions (*Left*), conventional uniform/coarse sampling fails to capture fine-grained visual evidence, whereas our spatial zoom-in highlights query-relevant regions and enables correct prediction. For event-level understanding questions (*Right*), prior approaches suffer from limited temporal coverage, while our temporal zoom-out expands long-range temporal context, allowing the model to identify the global temporal structure. Together, these results demonstrate that LENS adaptively allocates visual evidence across spatial and temporal dimensions to improve video reasoning.

to tens of millions of visual tokens when processed by models such as LLaVA-OneVision [15], far exceeding typical context limits. As a result, directly feeding full videos into the model is computationally prohibitive, necessitating mechanisms that enable MLLMs to efficiently handle video understanding tasks.

To address the context bottleneck, a natural strategy is to select only a small subset of frames from the video as visual inputs for downstream reasoning. The efficacy of long-form video understanding is therefore intrinsically tied to the quality of this selection; any omission of pivotal moments leaves the model with insufficient information to address the query. Despite this, many state-of-the-art MLLMs [2, 4, 15, 24, 52, 58] still rely on naive uniform sampling strategies that select frames at fixed intervals, which can miss task-critical content when relevant evidence is sparse or localized in time. This highlights the practical importance of *keyframe sampling*, where selecting informative and query-relevant frames is essential for effective long-form video reasoning. However, determining which frames are truly informative is inherently challenging, as relevant visual cues may be sparse, subtle, or distributed across distant segments of the video.

To tackle this problem, recent works have explored various keyframe sampling strategies aimed at identifying frames that are most informative for answering a given query. For example, BOLT [21] and AKS [31] use image–text matching models to estimate query–frame relevance scores, then apply inverse transform sampling or adaptive judge-and-split procedures to balance relevance with temporal diversity. In contrast, T* [44] utilizes object detection models to identify inferred cue objects and iteratively refine its frame selection process. However,

these approaches primarily operate at the frame level and treat frames as independent sampling units, focusing on selecting which frames to retain rather than how visual evidence should be allocated across spatial and temporal dimensions. Consequently, their fixed-granularity design limits flexibility under diverse query demands, as selected frames alone may not provide sufficient spatial precision or temporal coverage, as illustrated in Figure 1.

In this work, we introduce LENS, a novel framework for keyframe sampling that dynamically adjusts its focus between fine-grained spatial details and long-range temporal context. Our key insight is that different queries require different forms of visual evidence: object-centric questions rely on fine-grained local cues, whereas event-level reasoning depends on broader temporal context. Instead of selecting a static subset of frames, LENS dynamically allocates frame budgets to emphasize either detailed visual regions or wider temporal coverage according to the query. As illustrated in Figure 1, LENS enables multi-granular reasoning that more effectively aligns visual evidence with the demands of the task.

Specifically, LENS realizes this adaptive strategy through two complementary operations: spatial zoom-in and temporal zoom-out.[1] The spatial zoom-in module highlights query-relevant regions within selected frames to capture fine-grained visual cues, while the temporal zoom-out module aggregates neighboring frames to expand the temporal field of view and preserve contextual continuity. A lightweight query-aware controller dynamically determines how to allocate the frame budget between these two branches, allowing the model to flexibly prioritize local precision or global context depending on the task. Notably, LENS is training-free and plug-and-play, making it directly applicable to both open-source and proprietary MLLMs without additional fine-tuning.

We comprehensively evaluate our approach on multiple long-form video question answering benchmarks and demonstrate that LENS consistently outperforms state-of-the-art training-free sampling methods across frame budgets and backbones. Qualitative analyses further highlight LENS's ability to flexibly balance spatial precision and temporal coverage, showing that it dynamically allocates frame budgets between spatial zoom-ins and temporal zoom-outs based on the text query, prioritizing fine-grained visual regions for detail-oriented questions while expanding temporal context for queries that require broader event-level understanding.

Our contributions are summarized as follows:

- We propose LENS, a novel training-free framework that adaptively partitions a limited frame budget between spatial zoom-in and temporal zoom-out operations, enabling multi-granular reasoning that balances fine-grained visual details with long-range temporal context.
- We devise spatial zoom-in and temporal zoom-out mechanisms that refine local evidence through query-conditioned visual prompting and capture global video structure via graph-based reasoning with hyperframe aggregation.

[1] We use "zoom" as an analogy along two orthogonal axes rather than in the strict optical sense. *Spatial zoom-in* narrows the spatial field of view within a frame to amplify local detail, whereas *temporal zoom-out* widens the temporal field of view by aggregating neighboring frames.

- We validate the effectiveness of LENS through comprehensive experiments on three long-form video understanding benchmarks, demonstrating that it consistently outperforms state-of-the-art keyframe sampling approaches.

2 Related Work

Long-Form Video Understanding. Recent progress in multi-modal learning has enabled large language models to perceive videos by augmenting them with visual encoders, converting frames into visual tokens that can be processed alongside text for a wide range of video understanding tasks [1–3,9,10,33,61]. However, early approaches struggle to handle hour-long videos in a single pass, as the resulting token sequences quickly exceed the available context length [9,14,26,40,55]. To address this challenge, recent work has focused on the fundamental problem of scaling long-context MLLMs and improving their temporal reasoning capabilities [27,40]. For instance, LongVU [28] introduces spatiotemporal adaptive compression to reduce temporal/spatial redundancy and fit many frames into a fixed window, and LongVILA [7] provides a full-stack solution that extends context length and performs long-video supervised fine-tuning with system-level optimizations. Another line of work tackles long-form video understanding from an inference-time perspective by designing LLM-based agents equipped with external tools to iteratively explore videos and gather relevant evidence. Representative methods include VideoAgent [35], VideoLucy [62], and VideoTree [37], which dynamically select tools (*e.g.*, frame captioning, temporal navigation) to decompose complex queries into sequential reasoning steps. More recently, retrieval-augmented generation (RAG) techniques have been adopted to enhance LLMs by building external knowledge bases and retrieving relevant information to support long-context reasoning [22,29,41]. However, the heavy computational overhead introduced by iterative tool calling hinders their scalability and practical deployment.

Video Keyframe Sampling. Orthogonal to these lines of work, we focus on the problem of *video keyframe sampling*, which aims to select a compact, query-relevant subset of frames that maximizes useful evidence [5,6,13,32,42,46]. Early methods such as BOLT [21] and AKS [31] frame keyframe selection as a trade-off between relevance and temporal diversity, leveraging inverse transform sampling or adaptive judge-and-split procedures to balance prompt–frame relevance with coverage. T* [44] advances this line of work by introducing a temporal search mechanism that detects visual cues and progressively refines sampling, upsampling relevant temporal or spatial regions using object detection models. Further, Q-Frame [56] performs query-aware frame selection with multi-resolution adaptation, allocating higher resolution to the most informative frames while keeping the overall token budget fixed. However, these approaches operate at a single granularity, focusing on which frames to select rather than how evidence should be distributed across spatial and temporal scales. In contrast, we introduce spatial zoom-in and temporal zoom-out to enable multi-granular reasoning that adaptively balances spatial precision and temporal coverage according to the query.

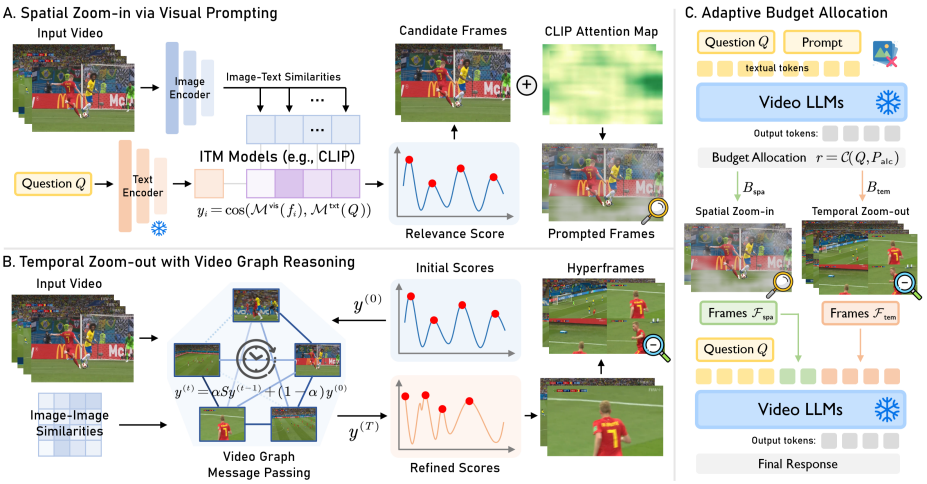

Fig. 2: Overview of LENS. (A) Spatial zoom-in selects query-relevant frames using image–text similarity and enhances fine-grained evidence via CLIP attention–based visual prompting. (B) Temporal zoom-out constructs a video graph to model inter-frame relationships and propagates relevance scores through message passing, enabling long-range reasoning and the selection of context-rich hyperframes. (C) An adaptive budget allocator dynamically distributes frame budgets between spatial and temporal branches, whose outputs are combined for final reasoning with a video LLM.

3 Method

In this work, we introduce LENS (Figure 2), a novel training-free framework that adaptively samples frames by performing spatial zoom-ins and temporal zoom-outs, enabling the model to reason across multi-granular scales to achieve both high-fidelity detail and comprehensive temporal coverage.

3.1 Preliminaries

MLLM-Based Video Understanding. Consider a video question answering task with an input video V consisting of T discrete frames $\{f_1, f_2, \cdots, f_T\}$ and an associated natural language query Q. To generate a linguistically coherent and visually grounded response, the MLLM employs a visual encoder $\mathcal{E}$ to extract latent features from each frame and a feature projector $\mathcal{P}$ to map these features into a sequence of visual tokens $\mathbf{x}_V = \mathcal{P}(\mathcal{E}(f_1), \mathcal{E}(f_2), \cdots, \mathcal{E}(f_T))$. These visual tokens are concatenated with the tokenized textual query $\mathbf{x}_T$ and fed into an LLM decoder for auto-regressive next-token generation. Formally, at each decoding step j, the next token y_j is sampled from the output probability distribution p_θ conditioned on the multimodal context and previously generated tokens $\mathbf{y}_{<j}$, represented as $y_j \sim p_\theta(y_j \mid \mathbf{x}_V, \mathbf{x}_T, \mathbf{y}_{<j})$.

Keyframe Sampling. Processing the entire sequence of T frames is often computationally prohibitive due to the quadratic complexity of self-attention in

the LLM decoder. To mitigate this, a sampling strategy $\mathcal{S}$ is typically employed to select a subset of k keyframes $\{f_{s_1}, f_{s_2}, \cdots, f_{s_k}\}$, where $k \ll T$. By reducing the temporal redundancy of the input video, keyframe sampling ensures that the visual tokens $\mathbf{x}_V$ remain within the LLM's effective context window while preserving the essential semantic content required for accurate video understanding.

3.2 Spatial Zoom-in via Visual Prompting

In long-form videos, query-relevant evidence is frequently sparse and localized, yet most approaches [21,31] rely on coarse frame-level representations that overlook such spatial nuances. We introduce a *pixel-level visual prompting* module that enhances spatial granularity by applying a query-conditioned attention mask to each frame. Instead of treating visual inputs uniformly, the module emphasizes semantically relevant and visually salient regions, effectively performing a *spatial zoom-in* that preserves visual context while capturing fine-grained cues.

Candidate Frame Selection. To identify a subset of query-relevant frames $\mathcal{F}'_{\mathsf{spa}}$ for subsequent visual prompting, we first extract a pool of candidate frames from the input video. We utilize a pre-trained image–text alignment model $\mathcal{M}$ (*e.g.*, CLIP [25] or BLIP [16]) to measure the relevance between each frame f_i and the query Q via cosine similarity, defined as

$$y_i = \cos(\mathcal{M}^{\mathsf{vis}}(f_i), \mathcal{M}^{\mathsf{txt}}(Q)), \quad i \in \{1, \cdots, T\}, \tag{1}$$

where $\mathcal{M}^{\mathsf{vis}}$ and $\mathcal{M}^{\mathsf{txt}}$ denote the visual and textual encoders, respectively. To prevent selected frames from clustering within a narrow temporal window, we implement a watershed algorithm [12] on the similarity curve to partition the video into distinct temporal segments. We then select the single highest-scoring frame from each segment as a representative. This ensures the candidate set $\mathcal{F}'_{\mathsf{spa}}$ spans the full video timeline rather than concentrating around a single event.

Query-Guided Visual Prompting. To highlight query-relevant regions within each frame of the candidate set $\mathcal{F}'_{\mathsf{spa}}$, we leverage cross-modal attention signals within a ViT-based CLIP [25] model $\mathcal{M}_{\mathsf{CLIP}}$ to derive spatial attribution maps that quantify the semantic relevance between the textual query and individual image patches, which enable us to localize fine-grained evidence while preserving the global visual context.

To decompose the global image–text similarity of each frame f into patch-level relevance scores, we formalize the final `[CLS]` representation as a linear aggregation of the value flows through the preceding attention and MLP layers following prior works [11]:

$$\mathcal{M}^{\mathsf{vis}}_{\mathsf{CLIP}}(f) = \phi\left([\mathbf{z}^0]_{\mathsf{cls}}\right) + \sum_{l=1}^{L} \phi\left(\left[\mathsf{MSA}^l(\mathbf{z}^{l-1})\right]_{\mathsf{cls}}\right) + \sum_{l=1}^{L} \phi\left(\left[\mathsf{MLP}^l(\tilde{\mathbf{z}}^l)\right]_{\mathsf{cls}}\right) \tag{2}$$

where L denotes the total number of transformer layers, $\mathbf{z}^{l-1}$ and $\tilde{\mathbf{z}}^l$ represent the token sequences prior to the l-th attention and MLP blocks respectively, and

where $\phi(\cdot)$ denotes the final linear projection (with normalization) that maps visual features into the joint embedding space. Prior work [19,45] shows that the final similarity is primarily governed by the deepest Multi-Head Self-Attention (MSA) layers, i.e., $l \in \{L', \ldots, L\}$, while the initial embedding $\mathbf{z}^0$ and MLP outputs act mainly as residual contributions.

We further decompose the outputs of the MSA blocks as a weighted summation over H attention heads and N input tokens:

$$\sum_{l=L'}^{L} \phi\left(\left[\mathsf{MSA}^l(\mathbf{z}^{l-1})\right]_{\mathsf{cls}}\right) = \sum_{l=L'}^{L}\sum_{h=1}^{H}\sum_{n=0}^{N} \phi\left(\alpha_n^{(l,h)} W^{(l,h)} z_n^{l-1}\right) = \sum_{l=L'}^{L}\sum_{h=1}^{H}\sum_{n=0}^{N} s^{(n,l,h)}, \tag{3}$$

where $\alpha_n^{(l,h)}$ represents the attention weight assigned to the n-th token by the h-th head at layer l relative to the `[CLS]` token. The term $\mathbf{V}_n^{(l,h)} = W_v^{(l,h)} z_n^{l-1}$ denotes the linear value projection of the n-th token. The term $s^{(n,l,h)}$ encapsulates the individual contribution of the n-th token to the final representation through a specific head and layer. By aggregating these contributions across all terminal layers and attention heads, we derive the final relevance score for each image patch $n \in \{1, \ldots, N\}$ as $\mathbf{c}_n = \sum_{l=L'}^{L}\sum_{h=1}^{H} s^{(n,l,h)}$. We then compute a query-conditioned relevance score by measuring alignment with the text embedding $r_n = \langle \mathbf{c}_n, \mathcal{M}_{\mathsf{CLIP}}^{\mathsf{txt}}(Q)\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the inner product. Intuitively, this score measures how strongly each patch contributes in the direction of the query embedding.

Saliency-Based Visual Prompting. Empirically, these query-conditioned relevance scores are most effective when the query Q specifies concrete objects or actions, as they can precisely localize regions aligned with explicit semantics. However, their discriminative capability weakens for more general queries, where relevance is not anchored to particular entities and instead depends on broader contextual cues. To mitigate this limitation, we introduce a complementary visual-only attribution that leverages the attention received from the `[CLS]` token in the final transformer layer, which has been shown to encode global contextual information [43,49]. Specifically, let $\mathbf{A}^{(L,h)} \in \mathbb{R}^{(N+1)\times(N+1)}$ denote the attention matrix of the h-th head in the final layer L. We compute the visual-only contribution score for a specific token n by aggregating the attention weights from `[CLS]` across heads:

$$\hat{r}_n = \frac{1}{H}\sum_{h=1}^{H} \mathbf{A}_{\mathsf{cls}\to n}^{(L,h)}, \quad n \in \{1, \ldots, N\}. \tag{4}$$

To ensure the final mask captures both explicit semantic targets and general foreground saliency cues, we combine the query-conditioned relevance r_n and the query-agnostic importance $\hat{r}_n$ using a soft OR fusion: $r_n^{\text{final}} = 1 - (1 - r_n)(1 - \hat{r}_n)$, which increases whenever either signal indicates high relevance.

The fused token scores are arranged into an attribution map on the token grid and resized to the original image resolution to obtain a raw heatmap. To reduce discretization artifacts from patch-level tokens, we apply a mean filter to produce smoother spatial transitions that better align with semantic boundaries.

This refined heatmap is then overlaid onto the original frames as an alpha mask, yielding a set of visually prompted frames $\mathcal{F}_{\mathsf{spa}}$, which effectively suppress irrelevant background noise and amplify query-relevant foreground regions.

3.3 Temporal Zoom-out with Video Graph Reasoning

Complementarily, long-form reasoning also requires understanding broader temporal context beyond isolated frames. To this end, we introduce a parallel branch with a temporal zoom-out mechanism that expands the temporal receptive field by aggregating information from a wider frame neighborhood. Instead of focusing on local evidence, this branch summarizes global temporal dynamics and contextual cues, enabling the model to capture long-range dependencies. By integrating fine-grained spatial evidence with coarse-grained temporal context, the framework achieves a more holistic understanding of video content.

Video Graph Reasoning. To explicitly model the temporal structure and semantic relationships among frames, we represent the input video as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node corresponds to a frame representation. We compute pairwise distances between frames using visual similarity measures such as CLIP image–image similarity [25] or SSIM [36]. Specifically, the distance between two frames is defined as $d(f_i, f_j) = (1 - \mathsf{Sim}(f_i, f_j))/2 \in [0, 1]$, where $\mathsf{Sim}(\cdot, \cdot)$ denotes the similarity score. We then sort all pairwise distances in ascending order and iteratively connect frame pairs following this order until the resulting graph becomes fully connected. Intuitively, this yields a sparse yet globally coherent graph that captures intrinsic video structure.

For all connected edges, we define the affinity matrix W using a Gaussian kernel to represent the pairwise similarity between frames:

$$W_{ij} = \begin{cases} \exp\left(-\frac{d^2(f_i, f_j)}{2\sigma^2}\right) & \text{if } i \neq j, \\ 0 & \text{if } i = j, \end{cases} \tag{5}$$

where $W_{ii} = 0$ effectively prevents self-loops. We then compute the symmetrically normalized adjacency matrix $S = D^{-1/2} W D^{-1/2}$, where D is the diagonal degree matrix with $D_{ii} = \sum_j W_{ij}$. Using this normalized structure, we perform iterative message passing until convergence according to the following update rule:

$$y^{(t)} = \alpha S y^{(t-1)} + (1 - \alpha) y^{(0)}, \tag{6}$$

where $y^{(0)} = y$ represents the initial image-text similarity scores between each frame and the text query Q in Equation (1), and $\alpha \in (0, 1)$ is a balancing hyperparameter that controls the trade-off between the local unary priors and the global graph structure. While Equation (6) admits a closed-form solution with $\mathcal{O}(T^3)$ complexity (T: total frames), we use the iterative form with $\mathcal{O}(K|\mathcal{E}|)$ cost ($|\mathcal{E}|$: edges, K: iterations), which is empirically negligible compared to MLLM processing time. Formally, since the symmetrically normalized matrix S has eigenvalues bounded within $[-1, 1]$, the message-passing sequence $\{y^{(t)}\}$ is a convergent Neumann series, which reaches a unique steady state. Concretely, $\{y^{(t)}\}$ converges to $y^* = (1-\alpha)(I - \alpha S)^{-1} y^{(0)}$ (see Zhou *et al.* [59] for a rigorous proof).

Intuitively, the refined scores obtained after the iterative updates produce a refined similarity landscape where the relevance of each frame is reinforced by the underlying video structure. Finally, we similarly apply watershed sampling over the refined scores to select another set of candidate frames $\mathcal{F}'_{\mathsf{tem}}$ that are globally representative of the video's semantics.

Hyperframe Aggregation. Given the selected candidate frames $\mathcal{F}'_{\mathsf{tem}}$, we construct *hyperframes* to enrich temporal context while keeping the token budget fixed. Specifically, for each candidate frame, we concatenate its neighboring $\tau - 1$ frames along the spatial axis to form a composite frame. The resulting hyperframe is then downsampled to a lower spatial resolution, trading spatial detail for broader temporal coverage, yielding the hyperframe set $\mathcal{F}_{\mathsf{tem}}$. This design enables the model to capture short-range motion and contextual cues around key moments without introducing additional frames, thereby improving temporal reasoning while maintaining computational efficiency.

3.4 Adaptive Keyframe Budget Allocation

The spatial and temporal branches described above provide complementary evidence: spatial zoom-in captures fine-grained local cues, while temporal zoom-out summarizes long-range context via video graph reasoning and hyperframes. However, different queries demand different levels of detail and contextual breadth. To reconcile this variability, we introduce a query-aware allocation mechanism that dynamically distributes a fixed frame budget B between the two branches.

Given a natural language query Q and an allocation prompt P_{alc}, we employ the MLLM as a lightweight controller $\mathcal{C}$ that predicts an allocation ratio[2]

$$r = \mathcal{C}(Q, P_{\mathsf{alc}}), \qquad B_{\mathsf{spa}} = \left\lfloor rB + \tfrac{1}{2} \right\rfloor, \qquad B_{\mathsf{tem}} = B - B_{\mathsf{spa}}, \tag{7}$$

where $r \in [0, 1]$ denotes the proportion of frames assigned to spatial zoom-in, and B_{spa} and B_{tem} specify the frame budgets for the spatial and temporal branches, respectively, i.e., $|\mathcal{F}_{\mathsf{spa}}| = B_{\mathsf{spa}}$ and $|\mathcal{F}_{\mathsf{tem}}| = B_{\mathsf{tem}}$. The final frame set is then formed by merging the two branches $\mathcal{F} = \mathsf{Sort}_t(\mathcal{F}_{\mathsf{spa}} \cup \mathcal{F}_{\mathsf{tem}})$, where $\mathsf{Sort}_t(\cdot)$ orders frames according to their original temporal indices to preserve chronological consistency. These frames are then provided to the MLLM as visual inputs for multi-modal reasoning.

4 Experiments

In this section, we validate the effectiveness of our keyframe sampling approach on various standard long video understanding benchmarks.

[2] See the Appendix for the prompt details.

Table 1: Evaluation results of VQA accuracy across different frame budgets on Video-MME, LongVideoBench (LVB), and MLVU. Missing entries are denoted by "–", indicating that results are either not reported or not applicable.

Method	Size	# Frames	Video-MME (w.o. sub.)				LVB	MLVU
			Short	Medium	Long	Overall		
Avg. Video Duration			*1.3min*	*9min*	*41min*	*17min*	*12min*	*12min*
MLLMs w/ Uniform Sampling								
Video-LLaVA [17]	7B	8	45.3	38.0	36.2	39.9	39.1	47.3
VideoLLaMA2 [9]	7B	16	56.0	45.4	42.1	47.9	-	-
LLaVA-NeXT-QW2 [18]	7B	8	58.0	47.0	43.4	49.5	-	-
LongVILA [7]	8B	128	60.2	48.2	38.8	49.2	-	-
LongVA [55]	7B	128	61.1	50.4	46.2	52.6	-	-
Video-XL [30]	7B	128/256	64.0	53.2	49.2	55.5	-	64.9
LLaVA-OneVision [15]	7B	*	64.0	53.2	49.2	58.2	56.3	64.7
LongVU [28]	7B	1fps	64.7	58.2	59.5	60.9	-	65.4
SF-LLaVA-1.5 [40]	7B	128	-	-	-	63.9	62.5	71.5
ViLAMP [8]	7B	1fps	-	-	57.8	67.5	61.2	-
Keye-VL-1.5 [34]	8B	64	81.2	70.7	67.1	73.0	66.0	-
Training-free Keyframe Sampling Approaches								
Qwen2.5-VL [1]	7B	8	60.9	51.4	47.4	53.3	53.3	52.8
+ BOLT [21]	7B	8	66.0	54.6	50.4	57.0	55.6	59.0
+ AKS [31]	7B	8	62.1	51.4	48.8	54.1	52.0	53.7
+ Q-Frame [56]	7B	8	66.7	58.7	51.3	58.7	56.6	58.2
+ LENS (**Ours**)	7B	8	**68.7**	**59.6**	**53.7**	**60.7**	**59.3**	**63.7**
Qwen2.5-VL [1]	7B	16	67.3	55.0	48.9	57.1	56.7	57.1
+ BOLT [21]	7B	16	71.7	57.7	49.7	59.7	58.0	64.5
+ AKS [31]	7B	16	66.4	56.9	48.6	57.3	55.7	57.8
+ Q-Frame [56]	7B	16	70.0	59.4	52.6	60.7	57.1	61.9
+ LENS (**Ours**)	7B	16	**72.3**	**62.1**	**54.8**	**63.1**	**59.7**	**65.9**
Qwen2.5-VL [1]	7B	32	72.6	59.0	51.8	61.1	58.4	59.4
+ BOLT [21]	7B	32	74.3	64.2	53.8	64.1	58.6	66.3
+ AKS [31]	7B	32	66.4	56.9	48.6	57.3	58.3	63.1
+ Q-Frame [56]	7B	32	71.8	62.4	52.8	62.3	58.7	64.6
+ LENS (**Ours**)	7B	32	**74.6**	**70.0**	**56.7**	**67.1**	**60.6**	**66.7**
LLaVA-OneVision [15]	7B	8	63.9	51.2	43.9	53.2	51.8	59.2
+ LENS (**Ours**)	7B	8	**66.2**	**60.1**	**49.7**	**58.7**	**57.5**	**66.3**
GPT-5-Mini	-	8	74.4	61.1	58.9	64.8	55.0	49.1
+ LENS (**Ours**)	-	8	**78.6**	**62.7**	**65.6**	**69.0**	**61.7**	**53.8**

4.1 Experimental Settings

Models and Benchmarks.To evaluate our approach, we sample keyframes and process them using a suite of state-of-the-art MLLMs. Specifically, we employ Qwen2.5-VL [1], LLaVA-OneVision [15], and GPT-5-Mini as our foundational models for video understanding. We conduct extensive experiments on three long-form video understanding benchmarks:

- **Video-MME [10]** is a widely used benchmark designed to evaluate MLLMs across various video lengths, requiring the integration of both short-term temporal cues and long-term semantic context;
- **LongVideoBench (LVB) [38]** evaluates a model's ability to perform complex reasoning and information retrieval over extended durations;

- **MLVU [60]** covers a diverse range of disciplines and tasks, providing a holistic assessment of a model's capacity to process long-sequence video data in specialized domains.

Collectively, these benchmarks feature videos ranging from several seconds to hours, providing a comprehensive evaluation of the effectiveness of our approach in selecting the most representative keyframes for long-form video understanding.

Baselines. As a simple baseline, we include results from a uniform sampling baseline, where frames are selected at equal time intervals. Furthermore, we compare our performance against three state-of-the-art keyframe selection methods: BOLT [21], AKS [31], and Q-Frame [56]. To ensure a fair comparison, we report the performance of these baselines based on our re-implementation.

Implementation Details. For all evaluations, we adhere to the default query formats and prompts specified by the respective benchmarks to ensure consistency. Following the evaluation protocol in prior work [31, 56], we configure our method to sample 8, 16, and 32 frames from the input video. For the temporal zoom-out module, we set $\sigma = 0.3$, $\alpha = 0.8$, and $\tau = 4$ as the default hyperparameters for hyperframe generation; these choices are supported by ablation studies reported in the Appendix. To prevent redundancy between the two branches, we explicitly constrain the temporal zoom-out module to avoid selecting frames already chosen by spatial zoom-in. All experiments are conducted on a server equipped with 8× 48GB NVIDIA RTX 6000 Ada GPUs.

4.2 Results and Discussions

Results on Qwen-2.5-VL. Table 1 presents a comprehensive comparison between LENS and existing training-free keyframe sampling strategies under varying frame budgets. Across all budgets (8, 16, and 32 frames), LENS consistently achieves the best performance, demonstrating the effectiveness of adaptive multi-granular keyframe selection. Under the strict 8-frame constraint, LENS attains an overall accuracy of 60.7% on Video-MME, surpassing the second-best baseline Q-Frame by 2.0%. The gains are consistent across short, medium, and long videos, indicating that even with limited visual evidence, the combination of spatial zoom-in and temporal zoom-out provides more informative frame selection than relevance-only or heuristic-based sampling. As the frame budget increases, the advantage of LENS becomes more pronounced. At 16 frames, LENS reaches 63.1% overall accuracy, improving over Q-Frame by 2.4% and showing notable gains on medium and long videos, where modeling long-range dependencies becomes increasingly important. With 32 frames, LENS achieves 67.1%, outperforming all baselines by a substantial margin. Overall, these results highlight that LENS is both budget-efficient and scalable, delivering consistent gains with limited frames while further amplifying improvements as more budgets becomes available.

LENS also demonstrates consistent improvements on LongVideoBench and MLVU. Under the 8-frame setting, LENS achieves 59.3% on LVB and 63.7% on MLVU, outperforming the strongest baseline Q-Frame by 2.7% and 5.5%, respectively. The gains remain stable as the frame budget increases, reaching 60.6%

Fig. 3: Qualitative comparison on two representative samples from the Video-MME benchmark [10]. The selected frames from LENS are compared against those obtained via uniform sampling and AKS [31]. Note that frames obtained via spatial zoom-in are denoted by green circles, while those from temporal zoom-out are represented by green hexagons.

on LVB and 66.7 on MLVU with 32 frames. These results further demonstrate that LENS consistently selects more informative frames and generalizes well across different long-form video benchmarks.

Results on Other MLLMs. To verify that LENS generalizes beyond Qwen2.5-VL, we apply it to two additional backbones under the 8-frame budget. On LLaVA-OneVision, LENS improves the overall accuracy on Video-MME from 53.2% to 58.7%, with consistent gains on short, medium, and long videos. Similar trends are observed on other benchmarks, where LENS increases performance on LongVideoBench from 51.8% to 57.5% and on MLVU from 59.2% to 66.3%. On the GPT-5-Mini backbone, LENS also yields clear improvements, raising Video-MME overall accuracy from 64.8% to 69.0% and notably enhancing long-video performance (58.9% vs. 65.6%). Meanwhile, LENS improves LongVideoBench from 55.0% to 61.7% and MLVU from 49.1% to 53.8%. These results suggest that LENS serves as a model-agnostic, training-free plug-in that consistently enhances long-form video reasoning by providing complementary spatial evidence and temporal context under tight frame budgets.

Qualitative Results. Figure 3 presents two representative comparisons of frame selection on Video-MME. In the first (football event recognition), uniform sam-

pling distributes frames evenly across the timeline but misses key discriminative cues, while AKS retrieves partially relevant moments yet fails to capture the decisive textual evidence at sufficient spatial resolution; consequently, both methods produce incorrect predictions. In contrast, LENS uses spatial zoom-in to emphasize fine-grained visual cues (e.g., the event-title region) and temporal zoom-out to preserve the overall match context, enabling the model to correctly identify the event as the *FIFA World Cup 2022 Final*. The second comparison (table tennis rally understanding) further underscores the importance of multi-granular reasoning. Uniform sampling and AKS attend to isolated moments of the rally and overlook the sustained interaction between players, resulting in incorrect interpretations of the play. By adaptively allocating frames, LENS captures both localized player actions and the temporal continuity of the rally, allowing the model to correctly infer that the sequence depicts *a series of consecutive successful hits*. Overall, these results show that LENS selects frames that are not only semantically relevant but also complementary across spatial and temporal scales, yielding more reliable reasoning than prior keyframe sampling strategies.
Additional Results. Please refer to Appendix A for more experimental results on EgoSchema [23] and NExT-QA [39], which consistently demonstrates that our LENS can generalize to diverse video domains and reasoning types.

4.3 More Discussions

Efficiency Analysis. Table 2 reports the total runtime and accuracy of different methods under an 8-frame budget, evaluated on all 2,700 Video-MME samples using eight NVIDIA RTX 6000 Ada GPUs. While LENS incurs a modest runtime increase over uniform sampling (11 h 50 min *vs*. 8 h 13 min), it achieves the highest accuracy of 60.7%, corresponding to a substantial gain of +7.4%. Compared to prior adaptive sampling approaches such as AKS and Q-Frame, LENS yields consistently larger improvements with only moderate additional computational overhead. This overhead primarily arises from the lightweight allocation controller, as well as the extra computation introduced by attention-based prompting in spatial zoom-in and graph-based reasoning in temporal zoom-out; however, these components remain computationally inexpensive in practice. Overall, these results confirm that LENS provides a strong balance between accuracy and efficiency, making it well-suited for real-world long-video applications.

Notably, LENS remains far more efficient than even recent LLM-agent and retrieval-augmented video reasoning methods. For instance, VideoRAG [22] and Vgent [29] each require over 50 hours on the same benchmark—more than 4× the runtime of LENS—underscoring its practicality as a lightweight solution for long-form video understanding.

Table 2: Efficiency comparison on Video-MME [10]. We report the total runtime, the achieved accuracy, and the performance gains compared to the uniform sampling baselines.

Method	Runtime	Accuracy	Gain
Qwen-2.5-VL [1]	8 h 13 min	53.3	-
AKS [31]	9 h 55 min	54.1	+0.8
Q-Frame [56]	10 h 30 min	58.7	+5.4
LENS (Ours)	11 h 50 min	**60.7**	**+7.4**

Table 3: Ablation study of LENS components on Video-MME (without subtitles) using Qwen2.5-VL under an 8-frame budget. We progressively add query-aware scoring (CLIP/BLIP), spatial zoom-in, and temporal zoom-out modules to analyze their individual and combined contributions. Note that B_{spa} and B_{tem} denote the numbers of frames allocated to spatial zoom-in and temporal zoom-out, respectively.

Method	# Frames		Video-MME (w.o. sub.)			
	B_{spa}	B_{tem}	Short	Medium	Long	Overall
Qwen2.5-VL [1]	-	-	60.9	51.4	47.4	53.3
+ CLIP Scores [25]	-	-	65.8	53.6	50.4	56.6
+ BLIP Scores [16]	-	-	65.9	54.2	50.8	57.0
+ Spatial Zoom-in	8	-	68.8	58.0	52.0	59.6
+ Temporal Zoom-out	-	8	68.5	57.5	51.8	59.2
+ Fixed Budgets	4	4	68.9	59.0	52.8	60.2
+ Adaptive Budgets	*	*	68.7	59.6	53.7	60.7

We also provide the average per-query runtime breakdown on Video-MME, where numbers in parentheses denote the cumulative performance after adding each component. LENS takes 15.8s in total: 11.0s for standard MLLM processing (53.3), 0.8s for CLIP/BLIP similarity computation (56.6), 1.4s for visual prompting (59.6), 0.5s for graph reasoning (60.2), and 2.1s for budget allocation (60.7).

Ablation Studies. Table 3 evaluates the contribution of each component in LENS under a strict 8-frame budget. Starting from the vanilla Qwen2.5-VL baseline, incorporating query-aware frame scoring already yields notable improvements, where CLIP-based selection increases overall accuracy from 53.3% to 56.6%, and BLIP further improves it to 57.0%. We then evaluate the impact of our spatial zoom-in and temporal zoom-out modules, which capture fine-grained details and long-range temporal context, respectively. Allocating all frames to a single scale (spatial zoom-in or temporal zoom-out) provides moderate gains, suggesting that relying solely on localized detail or coarse temporal coverage is insufficient for complex video reasoning. In contrast, combining both through a fixed budget split (4 spatial / 4 temporal) substantially boosts performance to 60.2 overall, demonstrating the complementary nature of the two mechanisms: spatial zoom-in improves fine-grained perception (*e.g.*, small objects or subtle actions), while temporal zoom-out enhances global event understanding and long-range dependencies. Finally, our adaptive budget allocation achieves the best performance (60.7% overall) and delivers the largest improvements on medium and long videos, where the optimal balance between detail and coverage varies more significantly across queries. This result highlights the importance of dynamically adjusting computational resources based on query complexity and video structure, enabling LENS to allocate frames where they are most informative.

Budget Allocation. We test whether LENS depends on the capacity of the module that allocates the frame budget between spatial zoom-in and temporal zoom-out. Replacing our default allocator with two stronger proprietary models (GPT-5-Mini and Gemini-2.5-Pro), while keeping Qwen-2.5-VL fixed as the QA backbone, yields only marginal additional gains on Video-MME (+0.4% and

+0.7%, respectively). This indicates that the improvements stem primarily from the multi-granular zoom-in/zoom-out formulation rather than from the strength of the allocator. Consistent with this, our ablations show LENS still delivers significant gains even without the controller, demonstrating that the framework is robust to the allocation mechanism and does not rely on large proprietary models.

5 Conclusion

We present LENS, a training-free, plug-and-play framework for adaptive keyframe sampling in long-form video understanding. LENS explicitly models the trade-off between fine-grained spatial evidence and long-range temporal context, allocating a limited frame budget through complementary spatial zoom-in and temporal zoom-out operations. This multi-granular strategy enables MLLMs to align visual evidence with the demands of each query, without modifying model weights or adding computational overhead. Across diverse long-form video question answering benchmarks, LENS consistently outperforms existing keyframe sampling strategies, underscoring the value of adaptive spatio-temporal allocation for efficient video reasoning. Qualitative analyses further show that LENS yields more interpretable evidence selection, focusing on discriminative regions when fine detail is decisive and expanding temporal coverage for event-level understanding.

Acknowledgments

This work has been funded in part by the ARL award W911QX-24-F-0049, DARPA award FA8750-23-2-1015, ONR award N00014-23-1-2840, ONR MURI grant N00014-25-1-2116, and USDA-NIFA award 2021-67021-35329.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Bi, J., Bellos, F., Guo, J., Li, Y., Huang, C., Tang, Y., Song, L., Liang, S., Zhang, Z., Corso, J.J., et al.: When to think and when to look: Uncertainty-guided lookback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113 (2026)
3. Bi, J., Guo, J., Liang, S., Sun, G., Song, L., Tang, Y., He, J., Wu, J., Vosoughi, A., Chen, C., et al.: Verify: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity. arXiv preprint arXiv:2503.11557 (2025)
4. Bi, J., Liang, S., Zhou, X., Liu, P., Guo, J., Tang, Y., Song, L., Huang, C., Vosoughi, A., Sun, G., et al.: Why reasoning matters? a survey of advancements in multimodal reasoning (v1). arXiv preprint arXiv:2504.03151 (2025)
5. Bi, J., Sun, G., Vosoughi, A., Chen, C., Xu, C.: Diagnosing visual reasoning: Challenges, insights, and a path forward. arXiv preprint arXiv:2510.20696 (2025)

6. Buch, S., Nagrani, A., Arnab, A., Schmid, C.: Flexible frame selection for efficient video reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29071–29082 (2025)
7. Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., He, Y., Yin, H., Molchanov, P., Kautz, J., Fan, L., Zhu, Y., Lu, Y., Han, S.: LongVILA: Scaling long-context visual language models for long videos. In: International Conference on Learning Representations (2025), https://openreview.net/forum?id=wCXAlfvCy6
8. Cheng, C., Guan, J., Wu, W., Yan, R.: Scaling video-language models to 10k frames via hierarchical differential distillation. In: International Conference on Machine Learning. pp. 9964–9981. PMLR (2025)
9. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
10. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24108–24118 (2025)
11. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting CLIP's image representation via text-based decomposition. In: International Conference on Learning Representations (2024), https://openreview.net/forum?id=5Ca9sSzuDp
12. Haralick, R.M., Sternberg, S.R., Zhuang, X.: Image analysis using mathematical morphology. IEEE Transactions on Pattern Analysis and Machine Intelligence (4), 532–550 (1987)
13. Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B., et al.: M-llm based video frame selection for efficient video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13702–13712 (2025)
14. Jin, P., Takanobu, R., Zhang, W., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13700–13710 (2024)
15. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=zKv8qULV6n
16. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
17. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 5971–5984 (2024)
18. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023)
20. Liu, S., Zhang, C.L., Zhao, C., Ghanem, B.: End-to-end temporal action detection with 1b parameters across 1000 frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18591–18601 (2024)

21. Liu, S., Zhao, C., Xu, T., Ghanem, B.: Bolt: Boost large vision-language model without training for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3318–3327 (2025)
22. Luo, Y., Zheng, X., Li, G., Yin, S., Lin, H., Fu, C., Huang, J., Ji, J., Chao, F., Luo, J., Ji, R.: Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In: Advances in Neural Information Processing Systems (2025), https://openreview.net/forum?id=QaZxGWlbgO
23. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
24. Pi, R., Han, T., Zhang, J., Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., Zhang, T.: Mllm-protector: Ensuring mllm's safety without hurting performance. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 16012–16027 (2024)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
26. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323 (2024)
27. Santos, S., Farinhas, A., McNamee, D.C., Martins, A.: inf-Video: A Training-Free Approach to Long Video Understanding via Continuous-Time Memory Consolidation. In: International Conference on Machine Learning. pp. 52877–52893. PMLR (2025)
28. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. In: International Conference on Machine Learning. pp. 54582–54599. PMLR (2025)
29. Shen, X., Zhang, W., Chen, J., Elhoseiny, M.: Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. In: Advances in Neural Information Processing Systems (2025), https://openreview.net/forum?id=5xPvWat3IX
30. Shu, Y., Liu, Z., Zhang, P., Qin, M., Zhou, J., Liang, Z., Huang, T., Zhao, B.: Video-xl: Extra-long vision language model for hour-scale video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26160–26169 (2025)
31. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29118–29128 (2025)
32. Tang, Y.Y., Shimada, D., Hua, H., Huang, C., Bi, J., Feris, R., Xu, C.: Video-r4: Reinforcing text-rich video reasoning with visual rumination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8314–8325 (2026)
33. Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., et al.: Video understanding with large language models: A survey. IEEE Transactions on Circuits and Systems for Video Technology (2025)
34. Team, K.K., Yang, B., Wen, B., Liu, C., Chu, C., Song, C., Rao, C., Yi, C., Li, D., Zang, D., et al.: Kwai keye-vl technical report. arXiv preprint arXiv:2507.01949 (2025)

35. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
37. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3272–3283 (2025)
38. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
39. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9777–9786 (2021)
40. Xu, M., Gao, M., Li, S., Lu, J., Gan, Z., Lai, Z., Cao, M., Kang, K., Yang, Y., Dehghan, A.: Slowfast-LLaVA-1.5: A family of token-efficient video large language models for long-form video understanding. In: Conference on Language Modeling (2025), https://openreview.net/forum?id=L7jS3peM3w
41. Xue, Z., Zhang, J., Xie, X., yuxuan cai, Liu, Y., Li, X., Tao, D.: AdavideoRAG: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In: Advances in Neural Information Processing Systems (2025), https://openreview.net/forum?id=FDAI0PY9Qp
42. Yan, S., Xiong, X., Nagrani, A., Arnab, A., Wang, Z., Ge, W., Ross, D., Schmid, C.: Unloc: A unified framework for video localization tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13623–13633 (2023)
43. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19792–19802 (2025)
44. Ye, J., Wang, Z., Sun, H., Chandrasegaran, K., Durante, Z., Eyzaguirre, C., Bisk, Y., Niebles, J.C., Adeli, E., Fei-Fei, L., et al.: Re-thinking temporal search for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8579–8591 (2025)
45. Yu, R., Yu, W., Wang, X.: Attention prompting on image for large vision-language models. In: European Conference on Computer Vision. pp. 251–268. Springer (2024)
46. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-chained image-language model for video localization and question answering. Advances in Neural Information Processing Systems **36**, 76749–76771 (2023)
47. Zhang, C., He, J., He, J., Sycara, K., Xie, Y.: Evolving contextual safety in multimodal large language models via inference-time self-reflective memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41182–41192 (2026)
48. Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 21715–21737 (2024)
49. Zhang, C., Ma, K., Fang, T., Yu, W., Zhang, H., Zhang, Z., Mi, H., Yu, D.: VScan: Rethinking visual token reduction for efficient large vision-language models.

Transactions on Machine Learning Research (2026), https://openreview.net/forum?id=KZYhyilFnt
50. Zhang, C., Stepputtis, S., Sycara, K., Xie, Y.: Dual prototype evolving for test-time generalization of vision-language models. Advances in Neural Information Processing Systems **37**, 32111–32136 (2024)
51. Zhang, C., Wan, Z., Kan, Z., Ma, M.Q., Stepputtis, S., Ramanan, D., Salakhutdinov, R., Morency, L.P., Sycara, K.P., Xie, Y.: Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=tTBXePRKSx
52. Zhang, J., Li, Y., Chen, H., Lu, H., Xue, L., Wang, B., Liu, H.: Spacenum: Revisiting spatial numerical understanding in vlms. arXiv preprint arXiv:2605.23898 (2026)
53. Zhang, J., Qian, C., Sun, H., Lu, H., Wang, D., Xue, L., Liu, H.: Progresslm: Towards progress reasoning in vision-language models. arXiv preprint arXiv:2601.15224 (2026)
54. Zhang, J., Yao, D., Pi, R., Liang, P.P., Fung, Y.R.: Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. pp. 7510–7545 (2025)
55. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=3ORAWQVGlx
56. Zhang, S., Yang, J., Yin, J., Luo, Z., Luan, J.: Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22056–22065 (2025)
57. Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wang, G., et al.: Instruction tuning for large language models: A survey. ACM Computing Surveys **58**(7), 1–36 (2026)
58. Zhang, Y., Wu, J., Li, W., Li, B., MA, Z., Liu, Z., Li, C.: LLaVA-video: Video instruction tuning with synthetic data. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=EElFGvt39K
59. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B.: Learning with local and global consistency. Advances in Neural Information Processing Systems **16**, 321–328 (2003)
60. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025)
61. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: International Conference on Learning Representations (2024), https://openreview.net/forum?id=1tZbq88f27
62. Zuo, J., Deng, Y., Kong, L., Yang, J., Jin, R., Zhang, Y., Sang, N., Pan, L., Liu, Z., Gao, C.: Videolucy: Deep memory backtracking for long video understanding. In: Advances in Neural Information Processing Systems (2025), https://openreview.net/forum?id=To7Rs2wsTd