STVFocus: Query-guided Spatio-Temporal Visual Focusing for Video LLMs

Taehun Kong, Minyoung Park, and Sangjun Ahn

AI Lab, CTO Division, LG Electronics, South Korea
{th.kong,minyoung5.park,sangjun.ahn}@lge.com

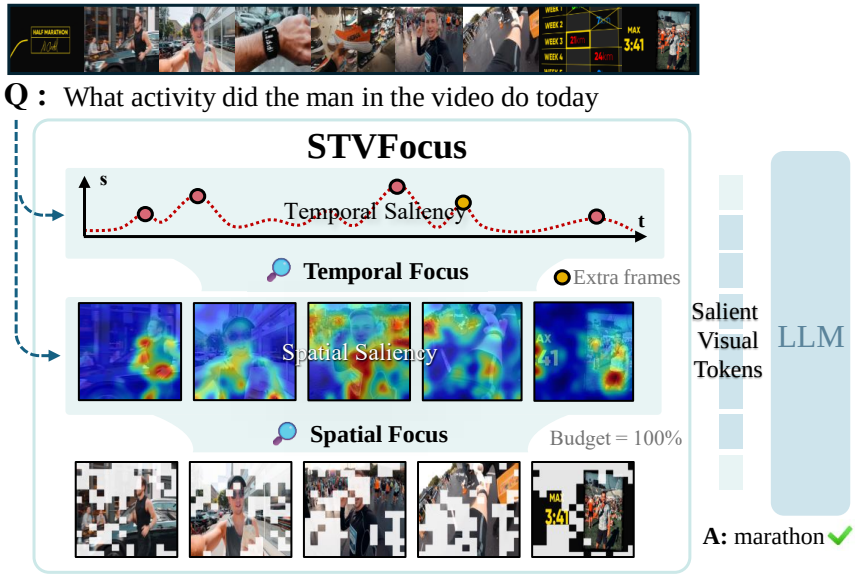


Fig. 1: Overview of the proposed STVFocus framework. Given a query and a video, STVFocus leverages query-aware spatio-temporal saliency to optimize visual context for Video LLMs. Temporal Focus identifies informative frames, and Spatial Focus highlights salient regions within each frame, while providing additional query-relevant temporal cues within the given budget. This focusing process produces a high information density visual context, enabling more accurate and coherent video understanding in LLMs.

Abstract. Video Large Language Models (Video LLMs) aim to understand and reason over dynamic visual content. However, videos inherently exhibit low task-relevant information density because only a small fraction of frames and regions provide meaningful cues. As a result, the useful signals are sparse across space and time. We present STVFocus, a training-free framework that unifies spatio-temporal focusing to capture salient contexts and build informative video representations. STVFocus consists of three components. First, Hierarchical Temporal Focus applies a coarse-to-fine frame selection strategy that captures global temporal structure while selectively emphasizing informative moments without excessive query bias. Second, Query-informed Spatial Focus uses spatial

saliency to highlight query-relevant intra-frame regions. Finally, Temporal Reallocation reinvests the saved spatial capacity into additional high-value frames, enabling richer temporal coverage. By jointly leveraging temporal and spatial saliency, STVFocus produces compact yet semantically enriched video representations. Experiments on VideoMME, LongVideoBench, and MLVU show consistent gains with meaningful margins, confirming the effectiveness of unified spatio-temporal focusing.

1 Introduction

Recent advances in Large Language Models (LLMs) have expanded their capabilities to visual modalities. Image-based multimodal LLMs [7, 28, 38] demonstrate strong visual reasoning by combining vision encoders with large language models. Building on this progress, Video LLMs [27, 30, 31, 35, 37, 46, 64, 65] extend these capabilities to video understanding. Unlike static images, videos exhibit both temporal continuity and spatial complexity, presenting unique challenges for reasoning over dynamic visual content. However, the inherently low information density of video data poses a fundamental limitation [44, 56]. Most video segments contain substantial amounts of task-irrelevant visual information that contribute little to meaningful interpretation. Temporally, consecutive frames frequently show almost the same content, providing only limited new information [58]. Spatially, static backgrounds and other uninformative regions make up a large portion of each frame [34]. This abundance of irrelevant information weakens meaningful signals and increases computational overhead, forcing models to process vast amounts of low-value visual input. Therefore, enhancing the representational quality of Video LLMs requires focusing on spatio-temporally salient regions that are important for effective understanding.

To address the prevalence of task-irrelevant information, recent studies have explored training-free approaches that selectively focus on salient content in the temporal or spatial domain. For temporal salient focusing, frame-level selection methods [22, 39, 53] identify informative frames by leveraging external multimodal embedding models [10, 29, 45, 61, 63]. While effective at highlighting temporally informative moments, these methods suffer from two key limitations: (1) the heavy external model is used only for temporal sampling and then discarded, leaving substantial computation underutilized, and (2) reliance on query-conditioned cues often introduces query bias that weakens short-video and global context understanding, limiting these methods to long-video settings. In contrast, spatial salient focusing has been far less explored. Most existing studies approach it indirectly through token compression [20,21,24,48], which largely prioritizes efficiency and fails to preserve spatial regions that are semantically important. A more recent attempt, Q-Frame [69] adjusts frame resolution based on temporal saliency. However, it still operates at the frame level and does not capture the fine-grained intra-frame cues essential for spatial understanding. To the best of our knowledge, no existing method explicitly guides spatial focusing within a frame according to query relevance. Overall, temporal and spatial

salient focusing have been explored separately. Yet existing approaches still fail to capture the synergy between temporal and spatial focusing needed to build information-dense video representations.

To overcome these limitations, we propose STVFocus, a training-free framework that unifies spatio-temporal focusing across frame and token levels by jointly leveraging temporal and spatial saliency, as shown in Fig. 1. STVFocus first performs **Hierarchical Temporal Focus**, a novel coarse-to-fine frame selection strategy guided by an external multimodal embedding model [45]. The coarse stage focuses on capturing the global temporal structure by selecting frames that outline the overall storyline and preserve contextual diversity. The fine stage then narrows the selection toward query-relevant moments, reflecting the tendency of human perception to process global structure before attending to local details [43]. This hierarchical process balances global coverage and temporal selectivity without being overly biased toward the query. In the spatial focusing stage, **Query-informed Spatial Focus** incorporates spatial saliency to identify query-relevant regions at the intra-frame level. It further extends the external multimodal embedding model beyond temporal focusing to enable fine-grained spatial focusing within each frame. By preserving important areas and suppressing redundant backgrounds, this spatial focusing mechanism refines spatial representations while maintaining contextual integrity. Finally, **Temporal Reallocation** expands temporal modeling capacity by directing attention toward additional frames of high temporal relevance. This redistribution enriches temporal coverage and results in a more information-dense video representation.

The main contributions of this work are summarized as follows:

- We present STVFocus, a training-free and model-agnostic framework for unified spatio-temporal focusing, jointly capturing temporal informativeness and spatial semantic relevance to build information-dense video representations.
- We introduce Hierarchical Temporal Focus, a coarse-to-fine frame selection strategy that organizes global temporal structure and progressively refines it toward query-relevant moments, producing a balanced set of diverse and informative frames.
- We propose Query-informed Spatial Focus and Temporal Reallocation for spatial-to-temporal refinement. Query-informed Spatial Focusing highlights semantically meaningful intra-frame regions, enabling Temporal Reallocation to use the resulting spatial capacity to extend temporal coverage.
- Extensive experiments demonstrate that STVFocus achieves consistent improvements across benchmarks, achieving 4.4% (+2.6) on VideoMME [19], 8.5% (+4.8) on LongVideoBench [57], and 8.7% (+5.5) on MLVU [73].

2 Related Works

2.1 Video Large Language Models

As Large Language Models (LLMs) expand from textual to visual domains, Video Large Language Models (Video LLMs) [11, 30, 35, 36, 41, 42, 60, 65] have

emerged to jointly model spatial and temporal information. Advances in this area stem from several complementary directions. Large-scale, high-quality video–text datasets drive substantial progress by providing stronger supervision for temporal reasoning [9, 12, 71]. Efforts to improve visual inputs explore dynamic-resolution modeling [4, 55, 64] and interleaved visual–textual tokens [1]. Another direction focuses on token-level modeling, refining visual features across spatial and temporal dimensions to enhance temporal understanding and representation efficiency [1, 26, 27, 40, 64]. LLaMA-VID [33] encodes each frame using only two tokens for compact yet expressive representations, while MovieChat [52] incorporates short-term and long-term memory tokens to capture temporal dependencies in long videos. Further work [32, 51] refines token representations inside the LLM to better support long-context video reasoning. However, these methods inherently depend on extensive training, which limits their scalability and practicality. This motivates the need for training-free, model-agnostic approaches that can efficiently enhance video understanding without additional optimization.

2.2 Temporal Context Adaptation in Video LLMs

Modern Video LLMs typically sample visual content at uniform temporal intervals, assuming all segments are equally informative. This often leads to missing salient video content, making retrieval somewhat dependent on chance. Recent work aims to better organize and structure the temporal context fed into Video LLMs, and these efforts can be broadly categorized into training-based and training-free methods. Training-based approaches perform temporal adaptation by teaching models to identify temporal segments that provide meaningful cues for downstream reasoning [2, 23, 62]. For instance, Frame-Voyager [62] learns temporal patterns that offer richer contextual grounding, while LongVU [50] reduces temporal redundancy by comparing visual feature similarity and adaptively maintains high-resolution visual information through alignment with text queries. However these approaches require additional training that increases computational cost. Complementary to these are training-free methods, which reorganize temporal content outside the Video-LLM pipeline using external models. BOLT [39] leverages multimodal embeddings [45, 63] with a CDF-based strategy to emphasize moments that are more strongly associated with the given query. In contrast, AKS [53] structures videos into temporal bins and applies a recursive judge-and-split process that balances relevance and temporal diversity.

Nonetheless, these methods rely on expensive external models that are used solely for temporal context adaptation and tend to be overly biased toward the query. Such query bias hinders global video understanding and limits these methods to long-video scenarios.

2.3 Visual Context Adaptation in Video LLMs

Although temporal context plays a crucial role in capturing the progression of events, effective video understanding also requires adapting the visual context

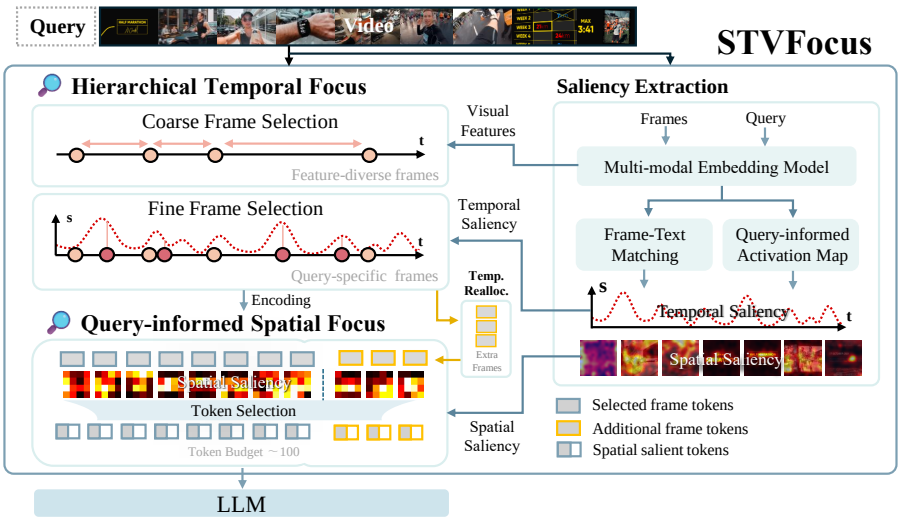


Fig. 2: The overall framework of STVFocus. STVFocus optimizes the visual context of Video LLMs by leveraging spatio-temporal saliency through three complementary modules. Hierarchical Temporal Focus (HTF) captures both globally representative moments and query-relevant fine-grained events, enhancing information density at the frame level. Query-informed Spatial Focus (QSF) directs attention toward salient spatial regions within each frame and refines intra-frame representation at the token level. Finally, Temporal Reallocation (TR) broadens temporal coverage by redistributing focus toward additional informative frames.

within the model itself. As part of this direction, token-level studies explore token compression as an approach to adapting visual context in Video LLMs.

Token compression, originally introduced in Vision Transformers [5, 13, 16], was later applied to Multimodal LLMs [8, 59, 75] and Video LLMs to simplify visual inputs or reduce redundancy. These approaches are typically categorized into Outer-LLM [25, 47, 49, 72] and Inner-LLM [20, 21, 70] strategies with hybrid variants [24, 48, 54]. However, most of them prioritize reducing token count rather than enhancing the semantic quality of visual representations. They also often encounter attention–semantic misalignment [68] or lack compatibility with efficient attention mechanisms such as Flash-Attention [14, 15]. A more recent line of work in visual context adaptation is represented by Q-Frame [69], which moves beyond simple token reduction toward reallocating visual capacity based on contextual importance. Q-Frame dynamically adjusts frame resolution according to temporal saliency, assigning higher spatial detail to informative moments while simplifying less relevant ones. However, its focus remains at the frame level, without addressing fine-grained intra-frame focusing.

In summary, existing visual context adaptation methods either focus on efficiency or overlook detailed spatial structures within frames. As a result, they fail to preserve important spatio-temporal and query-relevant cues.

3 STVFocus

3.1 Method Overview

The pipeline of our STVFocus framework is illustrated in Fig. 2. Guided by spatio-temporal saliency, STVFocus refines the visual context to highlight spatially meaningful regions and capture temporally rich and informative moments. The framework consists of four main stages. (1) Saliency Extraction: STVFocus extracts spatio-temporal saliency using an external multimodal embedding model [45], which measures the degree of relevance between the query and visual content both globally across frames and spatially within each frame. (2) Hierarchical Temporal Focusing: Using temporal saliency, a flexible iterative selection strategy progressively constructs an informative subset and extends it into a coarse-to-fine structure, enabling the selected frames to capture global context before focusing on fine-grained temporal details. (3) Query-informed Spatial Focusing: Guided by spatial saliency, it directs attention toward semantically meaningful tokens in the spatial domain, enabling the model to concentrate on task-relevant regions and strengthen spatial interpretability. (4) Temporal Reallocation: The capacity freed from suppressed spatial regions is reallocated to additional informative frames, enriching temporal diversity and strengthening contextual representation. Overall, STVFocus improves video representations at both frame and token levels, enabling the model to focus on spatio-temporally meaningful content for more comprehensive and context-aware interpretation.

3.2 Saliency Extraction

STVFocus extracts the query-related spatio-temporal saliency of each frame using a multimodal embedding model [45], decomposing it into temporal and spatial components to capture query relevance both across and within frames.

Temporal Saliency. The temporal saliency captures how much each frame contributes to the overall query relevance. Let the frame pool be defined as $F = \{I_1, \ldots, I_N\}$, where N denotes the size of the frame pool. Given a text query Q, its embedding $\mathbf{q} = f_T(Q)$ is obtained through the text encoder f_T. For each frame $I_i \in F$, we extract the visual embedding $\mathbf{v}_i = f_V(I_i)$ and compute its temporal saliency score $s_i \in \mathbb{R}$ as the cosine similarity between $\mathbf{q}$ and $\mathbf{v}_i$:

$$s_i = \frac{\mathbf{q} \cdot \mathbf{v}_i}{|\mathbf{q}||\mathbf{v}_i|}, \quad i = 1, \ldots, N. \tag{1}$$

The obtained scores are then normalized and sharpened using a control factor α:

$$\tilde{s}_i = \left(\frac{s_i - \min(s)}{\max(s) - \min(s)} \right)^{\alpha}, \quad i = 1, \ldots, N. \tag{2}$$

Spatial Saliency. For a given frame I_i, the visual feature $\mathbf{v}_i$ is used to compute a spatial saliency map $\mathbf{M}_i \in \mathbb{R}^{H \times W}$:

$$\mathbf{M}_i = g(\mathbf{v}_i, \mathbf{q}), \tag{3}$$

where $\mathbf{M}_i(h, w)$ indicates the relevance of location (h, w) in frame i to the query Q. In our work, the function $g(\cdot)$ is implemented using gradient-based explainability methods [6] to generate $\mathbf{M}_i$, which is subsequently smoothed with a Gaussian filter to reduce noise (i.e., $\mathbf{M}_i \leftarrow \text{Gaussian}(\mathbf{M}_i)$).

The Saliency Extraction stage produces spatio-temporal saliency signals that inform temporal and spatial focusing in the subsequent modules.

3.3 Hierarchical Temporal Focus

Based on the extracted temporal saliency, we introduce a Hierarchical Temporal Focus (HTF) module that selects a subset of informative frames by jointly considering query relevance and visual diversity. HTF is inspired by Farthest Point Sampling (FPS) [17] for diverse coverage of visual content, and extends it into a coarse-to-fine framework reflecting the hierarchical nature of human visual perception.

Iterative Selection Mechanism. Let $\mathbf{v}_i$ denote the visual feature of frame I_i and $\tilde{s}_i$ its query relevance score obtained from the temporal saliency computation. HTF maintains a selected frame set $S \subseteq F$, initialized with a seed frame $S \leftarrow \{I_1\}$. At each iteration, the next frame is chosen from $F \setminus S$ to balance diversity and query relevance:

$$I_{\text{next}} = \arg \max_{I_j \in F \setminus S} \Big[\beta \min_{I_m \in S} D(\mathbf{v}_j, \mathbf{v}_m) + (1 - \beta)\tilde{s}_j \Big], \tag{4}$$

where $D(\cdot, \cdot)$ denotes a distance metric (e.g., $1 - \cos(\mathbf{v}_j, \mathbf{v}_m)$), and $\beta \in [0, 1]$ controls the trade-off between diversity and query relevance. The selected set is iteratively updated as $S \leftarrow S \cup \{I_{\text{next}}\}$ until k frames are chosen. Because each frame is selected sequentially based on the previously chosen subset, HTF forms an ordered selection process that captures inter-frame dependencies. This flexible, order-aware mechanism naturally allows HTF to extend into a coarse-to-fine strategy.

Coarse-to-Fine Strategy. HTF is designed to emulate human visual perception, where we first grasp the global context and then shift attention to task-relevant details [43]. For example, when watching a video, people naturally identify the overall type (e.g., news or advertisement) before attending to specific elements (e.g., product price or features). Following this coarse-to-fine perceptual process, HTF adopts a strategy controlled by β. In the coarse phase ($\beta = 1$), frames are selected primarily for diversity, forming a base set F_{coarse} that captures the overall video context. Decreasing β shifts the selection toward the fine phase, where selection prioritizes query relevance under a diversity regularizer, resulting in a *query-aligned* subset F_{fine}. The final selection is defined as

$$F_{\text{down}} = F_{\text{coarse}} \cup F_{\text{fine}}, \quad |F_{\text{down}}| = k, \tag{5}$$

producing an information-rich set that preserves global coverage while concentrating on query-specific content.

3.4 Query-informed Spatial Focus

Building upon the temporally focused frames F_{down} selected by HTF, we further enhance spatial representation within each frame through **Query-informed Spatial Focus (QSF)**. While HTF determines when to focus, QSF determines where to focus by identifying spatial regions that carry the most meaningful and query-relevant information. To achieve this, we integrate intrinsic activation cues from visual token features with query-guided spatial saliency maps $\mathbf{M}_i$. This fusion enables fine-grained focusing on semantically important areas while deemphasizing redundant or background regions.

Given the encoded patch token features $\mathbf{X} \in \mathbb{R}^{n_t \times (H_p W_p) \times D}$, where $\mathbf{X}_{t,p} \in \mathbb{R}^D$ denotes the feature vector of patch token p in frame t, we first compute an activation-based attention score [16]:

$$\mathcal{F}(\mathbf{X}_{t,p}) = \sum_{d=1}^{D} \left|\mathbf{X}_{t,p,d}\right| \in \mathbb{R}^+. \tag{6}$$

These scores are normalized to form an activation-based attention map (AAM):

$$\text{AAM}(t,p) = \frac{\mathcal{F}(\mathbf{X}_{t,p}) - \min(\mathcal{F}(\mathbf{X}))}{\max(\mathcal{F}(\mathbf{X})) - \min(\mathcal{F}(\mathbf{X}))}, \tag{7}$$

which captures the intrinsic semantic importance of each token, independent of the query.

To incorporate query awareness, the spatial saliency map $\mathbf{M}_i$ of frame I_i is downsampled via adaptive average pooling to align with the patch-token resolution:

$$\tilde{\mathbf{M}}_i = \text{AdaptiveAvgPool2d}(\mathbf{M}_i) \in \mathbb{R}^{H_p \times W_p}. \tag{8}$$

We then fuse the AAM with query-guided saliency $\tilde{\mathbf{M}}_i$ to obtain a query-informed spatial focus map:

$$\mathbf{s}_{\text{focus}} = \gamma \tilde{\mathbf{M}}_i + (1-\gamma)\text{AAM}, \quad \gamma \in [0,1], \tag{9}$$

where γ balances query-driven relevance and intrinsic activation strength.

Finally, spatial tokens are emphasized according to their focus scores $\mathbf{s}_{\text{focus}}$. We retain the most salient spatial regions while suppressing less informative ones:

$$\mathbf{X}' = \text{TopK}_\tau\left(\mathbf{X}, \mathbf{s}_{\text{focus}}\right), \tag{10}$$

where τ denotes a spatial retention rate that selects the top τ fraction of tokens and controls the degree of spatial emphasis.

By combining temporal context from HTF with query-informed spatial cues and intrinsic activations, QSF enables fine-grained spatial focusing that highlights informative regions while suppressing background content. This process produces context-aware frame representations that complement temporal focusing for improved video understanding.

3.5 Temporal Reallocation

Spatial focusing suppresses less informative regions and reduces computational cost, leaving extra capacity for temporal enhancement. Instead of treating this saved capacity as the end of the process, STVFocus uses it to expand temporal coverage. It adopts a budget-aware temporal enrichment scheme that reinvests unused spatial tokens into additional frame context. This mechanism is intentionally simple yet practical: HTF already produces an ordered list of frames, allowing STVFocus to incorporate additional frames in a straightforward top-down manner. Because the total token budget remains fixed, the model maintains stable throughput while flexibly reallocating capacity between spatial and temporal dimensions.

To support this reallocation, we utilize the frame ordering from HTF, which reflects each frame's temporal importance. Let B_{total} denote the total token budget and τ the spatial retention rate after QSF, such that the remaining budget becomes $B_{\text{remain}} = B_{\text{total}}(1-\tau)$. Assuming each frame contributes b_{frame} tokens, the number of additional frames that can be incorporated is computed as:

$$k_{\text{add}} = \left\lfloor \frac{B_{\text{remain}}}{b_{\text{frame}}} \right\rfloor. \tag{11}$$

We then expand the fine-phase frame subset by adding the next k_{add} frames from the ordered list generated by HTF:

$$F_{\text{final}} = F_{\text{down}} \cup \{I_j \in F \setminus F_{\text{down}} \mid j \leq k_{\text{add}}\}. \tag{12}$$

4 Experiments

4.1 Experimental Settings

Benchmarks. We evaluate STVFocus on three widely used VQA benchmarks: Video-MME [19], LongVideoBench (LVB) [57], and MLVU [73]. (1) Video-MME contains 900 videos ranging from short clips to hour-long content across diverse domains, with 2.7K curated QA pairs. STVFocus is evaluated without subtitles. (2) LVB includes 3.7K videos and 6.6K QA pairs up to about an hour, focusing on long-context reasoning and task-relevant moment identification. STVFocus is evaluated on the *val* set. (3) MLVU consists of videos ranging from 3 minutes to 2 hours across multiple genres and nine task types, assessing a model's ability to comprehend complex video content.

Implementation Details. To validate STVFocus, we use publicly available LLaVA-OneVision-7B [27], VILA1.5-8B [36] and InternVL3-8B [74] as base models. For spatio-temporal saliency extraction, we adopt CLIP-L14 [45] as the multimodal embedding model. The temporal sharpness factor α is set to 2, and the spatial saliency maps $\mathbf{M}_i$ are smoothed with a Gaussian filter ($\sigma = 8$) to reduce noise. The frame pool F is constructed by sampling each video at 1 FPS. For the coarse-to-fine selection in HTF, we allocate half of the total frames to the coarse phase and use the remaining half in the fine phase. In the fine-phase

Model	Size	Training-free	VideoMME				LVB	MLVU
			Overall	Short	Medium	Long		
Video-LLaVA [35]	7B	✗	39.9	45.3	38.0	36.2	39.1	47.3
VideoChat2 [31]	7B	✗	39.5	48.3	37.0	33.2	-	44.5
Chat-UniVi-V1.5 [26]	7B	✗	40.6	45.7	40.3	35.8	-	-
Qwen-VL [3]	7B	✗	41.1	46.9	38.7	37.8	-	-
LongVU [50]	7B	✗	60.9	64.7	58.2	59.5	-	65.4
LongVA [67]	7B	✗	52.6	61.1	50.4	46.2	-	-
Video-XL [51]	7B	✗	55.5	64.0	53.2	49.2	-	64.9
Video-CCAM [18]	9B	✗	50.3	61.9	49.2	39.6	-	58.5
LLaVA-OneVision [27]	7B	✗	58.5	70.3	56.6	48.8	56.4	63.2
+ AKS* [53]	7B	✓	59.6	71.1	57.4	50.1	59.3	-
+ BOLT [39]	7B	✓	59.9	70.1	60.0	49.6	59.6	66.8
+ **STVFocus**	7B	✓	**61.1**	**71.9**	**61.0**	**50.4**	**61.2**	**68.7**
VILA1.5 [36]	8B	✗	47.5	57.8	44.3	40.3	47.1	46.3
+ Frame-Voyager [62]	8B	✗	50.5	60.3	47.3	43.9	-	49.8
+ Q-frame [69]	8B	✓	50.7	59.8	48.0	44.2	51.6	**54.4**
+ **STVFocus**	8B	✓	**52.6**	**61.8**	**50.8**	**45.2**	**52.5**	50.9
InternVL3 [74]	8B	✗	64.5	75.7	64.1	53.8	58.5	67.8
+ **STVFocus**	8B	✓	**67.0**	**77.6**	**68.2**	**55.3**	**61.5**	**74.8**

Table 1: Performance comparison of state-of-the-art models across benchmarks. Gray rows indicate the baseline models without additional modules. Best results are highlighted in bold. The asterisk (*) denotes the reproduced performance of AKS [53] on the VideoMME benchmark. LLaVA-OneVision [27] and InternVL3 [74] use 32 frames, while VILA1.5 [36] uses 8 frames. STVFocus consistently achieves significant performance improvements with clear margins across different baseline models and benchmarks.

frame selection, the diversity–relevance balancing parameter β is set to 0.5. The token-level fusion parameter γ is fixed at 0.5 for all models. Finally, for QSF, the spatial focusing threshold τ is set to 80%. All experiments are conducted on four H100 GPUs using the LMMs-Eval framework [66], with a single set of hyperparameters across all datasets and base models.

4.2 Main Results

Tab. 1 compares STVFocus with other training-free methods that improve video understanding across multiple Video LLMs. When LLaVA-OneVision [27] is used as the base model, STVFocus consistently outperforms all previous methods on three benchmarks. On VideoMME, it improves performance from 58.5 to 61.1 (+2.6) and shows consistent gains across videos of different lengths, owing to the Coarse-to-Fine sampling strategy in HTF, which ensures balanced performance across videos of different lengths. Even when applied alone, HTF achieves higher accuracy than frame selection methods such as AKS [53] and BOLT [39] (see supplementary material for detailed results). In addition, STVFocus achieves substantial improvements on LVB and MLVU, increasing from 56.4 to 61.2 (+4.8)

HTF		QSF	Temp. Realloc.	Resource		Video MME	LVB
Coarse	Fine			# Frames	Token Budget		
Baseline				32	100%	58.5	56.4
✓				32	100%	59.2	56.8
	✓			32	100%	60.2	60.5
✓	✓			32	100%	60.8	60.6
✓	✓	✓		32	80%	60.8	60.7
✓	✓	✓	✓	32 + 8	100%	**61.1**	**61.2**

Table 2: Ablation studies of each component. Each component contributes to performance gains, and using all components together yields a notable improvement without exceeding the token budget.

Model	VideoMME			
	Overall	Short	Medium	Long
InternVL3-8B^{8frames}	59.8	69.1	60.1	50.1
+ **STVFocus**^{8frames}	**63.3**	**75.1**	**60.4**	**54.4**
InternVL3-8B^{16frames}	61.6	74.0	59.4	51.3
+ **STVFocus**^{16frames}	**64.5**	**75.9**	**63.7**	**53.9**
InternVL3-8B^{32frames}	64.5	75.7	64.1	53.8
+ **STVFocus**^{32frames}	**67.0**	**77.6**	**68.2**	**55.3**

Table 3: Scalability across frame budgets on VideoMME. STVFocus consistently improves InternVL3-8B [74] from 8 to 32 frames. For STVFocus, superscripts indicate frame-based total token budgets, not the raw number of sampled frames.

and 63.2 to 68.7 (+5.5), respectively—the largest performance margin among all compared methods.

A similar trend appears with VILA1.5, where STVFocus surpasses the previous state-of-the-art Q-Frame [69] by +1.9 on VideoMME, maintaining stable performance across various video lengths. Notably, on VideoMME, STVFocus surpasses the training-based method Frame-Voyager [62] by a considerably larger margin. While Q-Frame shows only a 0.2 difference from Frame-Voyager, STVFocus achieves a much wider +2.1 improvement, highlighting its strong effectiveness without additional training. Although Q-Frame shows higher performance on MLVU, STVFocus demonstrates greater overall improvements on LVB and VideoMME, reflecting its robustness across benchmarks.

Even on a recent and strong baseline like InternVL3, STVFocus improves VideoMME from 64.5 to 67.0 (+2.5), with consistent gains on LVB and a large +7.0 improvement on MLVU.

4.3 Ablation Studies and Analyses

Contribution of Each Component. Tab. 2 presents the results of each proposed component. To demonstrate the generality of our approach, we conducted ablation studies on two benchmarks, VideoMME and LVB. Even without query relevance, performing coarse frame selection based solely on visual features improves performance by +0.7 over the baseline. This indicates that considering global video context with diverse visual cues is more effective than uniformly sampling frames over time. Using only the fine frame selection module yields a

QSF		Token Budget	Video MME
AAM	Sal.		
		100%	58.5
✓		50%	59.1
	✓	50%	58.4
✓	✓	50%	**59.4**

(a) Effect of AAM and Spatial Saliency.

QSF γ	Video MME
0.3	58.8
0.5	**59.4**
0.6	59.2
0.7	58.6
Mult.	59.1

(b) Effect of γ in QSF.

Table 4: Ablation of QSF components and token-level fusion parameter γ. Results are obtained by applying QSF only to the baseline model, without HTF or temporal reallocation.

QSF τ	Video MME
90%	60.5
80%	**61.1**
60%	60.7
50%	60.9

(a) Effect of τ in QSF.

HTF β	Video MME
0.3	59.0
0.4	59.9
0.5	**60.2**
0.6	59.9
0.7	59.7

(b) Effect of β in HTF.

Table 5: Ablation of QSF focusing intensity threshold τ and HTF Diversity–Relevance Balancing Factor β.

+1.7 improvement. The coarse stage contributes more to short videos, while the fine stage brings larger gains for longer ones. When adopting the Coarse-to-Fine strategy, performance further improves by +2.3, validating the effectiveness of the hierarchical approach that first captures the overall video context and then focuses on fine-grained details. After applying QSF, the model achieves comparable or even higher performance while using only 80% of the tokens, demonstrating efficient token utilization. Finally, incorporating Temporal Reallocation further boosts performance by enhancing the temporal richness of the video representation.

Scalability Across Frame Budgets. Tab. 3 evaluates STVFocus with InternVL3-8B across frame budgets from 8 to 32. STVFocus delivers substantial improvements even under the 8-frame setting, demonstrating its ability to extract essential temporal and spatial cues from highly constrained inputs. As the number of frames increases, the performance gap remains consistent. This indicates that the gains arise from better capturing salient information while reducing redundancy, rather than simply leveraging more frames. Overall, the results confirm that STVFocus consistently improves performance across diverse temporal budgets.

Effect of Spatial Focusing. In QSF, spatial focusing is achieved by fusing two complementary cues: the intrinsic semantic importance derived from visual token features through AAM and the spatial importance estimated from Spatial Saliency. We first validate the individual contributions of AAM and Spatial Saliency, as shown in Tab. 4a. With a fixed focusing intensity of 50%, ap-

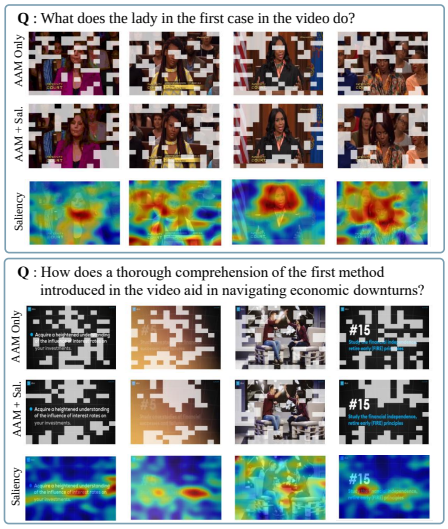
plying only AAM improves performance by +0.6, indicating that emphasizing semantically important tokens helps retain critical information. When relying solely on spatial saliency, the model attends to only half of the spatial areas yet achieves nearly the same performance as the baseline, suggesting that visual tokens exhibit inherent spatial correlation. When AAM and Spatial Saliency are fused, the model jointly considers both intrinsic and spatial importance, achieving 59.4, a +0.9 improvement over the baseline. This result demonstrates that spatial saliency effectively guides focusing, leading to the most balanced and informative spatial representation. As shown in Tab. 4b, we also evaluate the token-level fusion parameter γ, which controls the balance between AAM and Spatial Saliency in QSF. The best performance is achieved with $\gamma = 0.5$, indicating balanced fusion of both cues. Building on this, we further analyze how QSF interacts with Temporal Reallocation within the STVFocus framework. In this process, a trade-off arises between spatial compactness and temporal richness depending on the focusing intensity threshold τ. As shown in Tab. 5a, When evaluating performance across different τ values, the model performed best at $\tau = 80\%$. Unlike prior compression-oriented approaches that mainly target efficiency, our method aims to enhance representational quality, and we do not explore lower focusing levels. We also observed that applying dynamic thresholds to adaptively vary the number of focused regions or frames degraded the LLM's temporal understanding. Therefore, we adopt a fixed focusing threshold to maintain consistent and stable performance across videos.

Effect of the Diversity–Relevance Balancing Factor In the fine phase of HTF, frames are selected by jointly balancing query relevance and frame diversity, controlled by the hyperparameter β. A larger β emphasizes diversity, while a smaller β favors query relevance. When $\beta = 0$, the module operates as a top-k selector based solely on temporal saliency. As shown in Tab. 5b, we experimented with β values ranging from 0.3 to 0.7 on LLaVA-OneVision [27], and the best performance was achieved at $\beta = 0.5$.

Quality of Visual Context. As illustrated in Fig. 3a, the visual context optimized by STVFocus concentrates on spatio-temporally salient regions, providing richer temporal cues for reasoning. The selected coarse frames are both semantically and temporally diverse, allowing the model to first grasp the overall type and context of the video (e.g., movie, documentary) and then focus on the segments most relevant to the query. Spatially, STVFocus effectively removes irrelevant backgrounds and focuses on the subject of the scene, whereas the visual context without STVFocus often includes redundant or irrelevant content, with substantial background information remaining. As a result, STVFocus produces a visual context with high spatio-temporal information density, supplying the LLM with only the most essential cues for understanding.

Furthermore, as shown in Fig. 3b, spatial saliency serves as an effective guidance signal that preserves tokens in critical regions rather than suppressing them. In the top example of Fig. 3b, when using AAM only, the model tends to suppress tokens around the person's face and torso. However, when saliency is incorporated these regions are retained, preventing the loss of crucial semantic

(a) Visual context with and without STVFocus. STVFocus focuses on query-relevant regions while preserving temporal coverage.



(b) Qualitative results of QSF. Fusing spatial saliency improves spatial focus compared to AAM alone.

Fig. 3: Qualitative results of STVFocus and QSF.

information. In the bottom example, using AAM only results in the suppression of important textual regions, while incorporating saliency prevents such suppression and leads to improved performance in OCR-related tasks.

This reallocation strategically utilizes the preserved spatial capacity to incorporate additional temporally informative frames, thereby providing finer temporal granularity and contextual continuity. As a result, the final visual context retains only the spatio-temporally essential content while further enriching temporal cues. This produces a focused and information-dense representation that is fused with text tokens for joint decoding within the LLM.

5 Conclusions

We introduce the STVFocus framework that optimizes the visual context of Video LLMs by focusing on spatio-temporally salient information. STVFocus enhances the information density of video representations from the frame to the token level through saliency-guided focusing. The framework consists of three key components: Hierarchical Temporal Focus (HTF) for focusing on global context and fine-grained temporal details, Query-informed Spatial Focus (QSF) for attending to meaningful regions, and Temporal Reallocation (TR) that enriches temporal cues for coherent understanding. Evaluated on three Video LLMs and three benchmarks, STVFocus consistently achieves strong and reliable performance gains. It provides an effective training-free and model-agnostic approach for improving video understanding.

References

1. Ataallah, K., Shen, X., Abdelrahman, E., Sleiman, E., Zhu, D., Ding, J., Elhoseiny, M.: Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. arXiv preprint arXiv:2404.03413 (2024)
2. Ataallah, K., Shen, X., Abdelrahman, E., Sleiman, E., Zhuge, M., Ding, J., Zhu, D., Schmidhuber, J., Elhoseiny, M.: Goldfish: Vision-language understanding of arbitrarily long videos. In: European Conference on Computer Vision. pp. 251–267. Springer (2024)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
6. Bousselham, W., Boggust, A., Chaybouti, S., Strobelt, H., Kuehne, H.: Legrad: An explainability method for vision transformers via feature formation sensitivity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20336–20345 (2025)
7. Chen, G.H., Chen, S., Zhang, R., Chen, J., Wu, X., Zhang, Z., Chen, Z., Li, J., Wan, X., Wang, B.: Allava: Harnessing gpt4v-synthesized data for lite vision-language models. arXiv preprint arXiv:2402.11684 (2024)
8. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: 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. pp. 19–35. Springer (2024)
9. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
10. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16901–16911 (2024)
11. 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)
12. Cho, J.H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Rasheed, H., Sun, P., Huang, P.Y., Bolya, D., Jain, S., Martin, M., Wang, H., Ravi, N., Jain, S., Stark, T., Moon, S., Damavandi, B., Lee, V., Westbury, A., Khan, S., Krähenbühl, P., Dollár, P., Torresani, L., Grauman, K., Feichtenhofer, C.: Perceptionlm: Open-access data and models for detailed visual understanding. arXiv:2504.13180 (2025)
13. Choi, J., Lee, S., Chu, J., Choi, M., Kim, H.J.: vid-tldr: Training free token merging for light-weight video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18771–18781 (2024)
14. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 (2023)

15. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems **35**, 16344–16359 (2022)
16. Ding, S., Zhao, P., Zhang, X., Qian, R., Xiong, H., Tian, Q.: Prune spatio-temporal tokens by semantic-aware temporal accumulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16945–16956 (2023)
17. Eldar, Y., Lindenbaum, M., Porat, M., Zeevi, Y.Y.: The farthest point strategy for progressive image sampling. IEEE transactions on image processing **6**(9), 1305–1315 (1997)
18. Fei, J., Li, D., Deng, Z., Wang, Z., Liu, G., Wang, H.: Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. arXiv preprint arXiv:2408.14023 (2024)
19. 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 Computer Vision and Pattern Recognition Conference. pp. 24108–24118 (2025)
20. Fu, Q., Cho, M., Merth, T., Mehta, S., Rastegari, M., Najibi, M.: Lazyllm: Dynamic token pruning for efficient long context llm inference. arXiv preprint arXiv:2407.14057 (2024)
21. Fu, T., Liu, T., Han, Q., Dai, G., Yan, S., Yang, H., Ning, X., Wang, Y.: Framefusion: Combining similarity and importance for video token reduction on large vision language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22654–22663 (2025)
22. Guo, W., Chen, Z., Wang, S., He, J., Xu, Y., Ye, J., Sun, Y., Xiong, H.: Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. arXiv preprint arXiv:2503.13139 (2025)
23. 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 Computer Vision and Pattern Recognition Conference. pp. 13702–13712 (2025)
24. Huang, X., Zhou, H., Han, K.: Prunevid: Visual token pruning for efficient video large language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 19959–19973 (2025)
25. Hyun, J., Hwang, S., Han, S.H., Kim, T., Lee, I., Wee, D., Lee, J.Y., Kim, S.J., Shim, M.: Multi-granular spatio-temporal token merging for training-free acceleration of video llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23990–24000 (2025)
26. 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)
27. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
29. 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)

30. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
31. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
32. Li, X., Wang, Y., Yu, J., Zeng, X., Zhu, Y., Huang, H., Gao, J., Li, K., He, Y., Wang, C., et al.: Videochat-flash: Hierarchical compression for long-context video modeling. arXiv preprint arXiv:2501.00574 (2024)
33. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)
34. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. arXiv preprint arXiv:2202.07800 (2022)
35. 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 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5971–5984 (2024)
36. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pretraining for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
37. Lin, K., Ahmed, F., Li, L., Lin, C.C., Azarnasab, E., Yang, Z., Wang, J., Liang, L., Liu, Z., Lu, Y., et al.: Mm-vid: Advancing video understanding with gpt-4v (ision). arXiv preprint arXiv:2310.19773 (2023)
38. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
39. 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 Computer Vision and Pattern Recognition Conference. pp. 3318–3327 (2025)
40. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4122–4134 (2025)
41. Luo, R., Zhao, Z., Yang, M., Dong, J., Li, D., Lu, P., Wang, T., Hu, L., Qiu, M., Wei, Z.: Valley: Video assistant with large language model enhanced ability. arXiv preprint arXiv:2306.07207 (2023)
42. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
43. Navon, D.: Forest before trees: The precedence of global features in visual perception. Cognitive psychology **9**(3), 353–383 (1977)
44. Pan, B., Panda, R., Fosco, C., Lin, C.C., Andonian, A., Meng, Y., Saenko, K., Oliva, A., Feris, R.: Va-red^2: Video adaptive redundancy reduction. arXiv preprint arXiv:2102.07887 (2021)
45. 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)

46. 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)
47. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22857–22867 (2025)
48. Shao, K., Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: Holitom: Holistic token merging for fast video large language models. arXiv preprint arXiv:2505.21334 (2025)
49. Shen, L., Gong, G., He, T., Zhang, Y., Liu, P., Zhao, S., Ding, G.: Fastvid: Dynamic density pruning for fast video large language models. arXiv preprint arXiv:2503.11187 (2025)
50. 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. arXiv preprint arXiv:2410.17434 (2024)
51. 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 Computer Vision and Pattern Recognition Conference. pp. 26160–26169 (2025)
52. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., 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. pp. 18221–18232 (2024)
53. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29118–29128 (2025)
54. Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: Dycoke: Dynamic compression of tokens for fast video large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18992–19001 (2025)
55. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
56. Wang, Y., Chen, Z., Jiang, H., Song, S., Han, Y., Huang, G.: Adaptive focus for efficient video recognition. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 16249–16258 (2021)
57. 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)
58. Wu, X., Gao, C., Lin, Z., Wang, Z., Han, J., Hu, S.: Rap: redundancy-aware video-language pre-training for text-video retrieval. arXiv preprint arXiv:2210.06881 (2022)
59. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al.: Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. arXiv preprint arXiv:2410.17247 (2024)
60. Xu, L., Zhao, Y., Zhou, D., Lin, Z., Ng, S.K., Feng, J.: Pllava: Parameter-free llava extension from images to videos for video dense captioning. arXiv preprint arXiv:2404.16994 (2024)
61. 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)

62. Yu, S., Jin, C., Wang, H., Chen, Z., Jin, S., Zuo, Z., Xu, X., Sun, Z., Zhang, B., Wu, J., et al.: Frame-voyager: Learning to query frames for video large language models. In: International Conference on Learning Representations. vol. 2025, pp. 84154–84179 (2025)
63. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
64. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
65. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553 (2023)
66. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 881–916 (2025)
67. 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. arXiv preprint arXiv:2406.16852 (2024)
68. Zhang, Q., Cheng, A., Lu, M., Zhuo, Z., Wang, M., Cao, J., Guo, S., She, Q., Zhang, S.: [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. arXiv e-prints pp. arXiv–2412 (2024)
69. Zhang, S., Yang, J., Yin, J., Luo, Z., Luan, J.: Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. arXiv preprint arXiv:2506.22139 (2025)
70. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., et al.: Sparsevlm: Visual token sparsification for efficient vision-language model inference. arXiv preprint arXiv:2410.04417 (2024)
71. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
72. Zhang, Y., Lu, Y., Wang, T., Rao, F., Yang, Y., Zhu, L.: Flexselect: Flexible token selection for efficient long video understanding. arXiv preprint arXiv:2506.00993 (2025)
73. 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 Computer Vision and Pattern Recognition Conference. pp. 13691–13701 (2025)
74. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
75. Zou, X., Lu, D., Wang, Y., Yan, Y., Lyu, Y., Zheng, X., Zhang, L., Hu, X.: Don't just chase" highlighted tokens" in mllms: Revisiting visual holistic context retention. arXiv preprint arXiv:2510.02912 (2025)