EventSTU: Event-Guided Efficient Spatio-Temporal Understanding for Video-LLMs

Wenhao Xu[1] Xin Dong[1] Yue Li[1] Haoyuan Shi[1] Yueyi Zhang[2] Zhiwei Xiong[1]✉

[1] University of Science and Technology of China, Hefei, China
[2] Midea Group, Shanghai, China
wh-xu@mail.ustc.edu.cn, zwxiong@ustc.edu.cn
https://wh-xu1.github.io/EventSTU.github.io

Abstract. Video large language models have demonstrated strong video understanding capabilities, but suffer from high inference costs due to the massive number of tokens in long videos. Inspired by efficient biological vision systems, we propose EventSTU, a training-free framework guided by bio-inspired event cameras to significantly reduce redundant tokens, thereby enabling efficient spatio-temporal understanding. In the temporal domain, we design a coarse-to-fine keyframe sampling algorithm that exploits the change-triggered property of event cameras to eliminate redundant frames. In the spatial domain, we design an adaptive token pruning algorithm that leverages the visual saliency of events as a zero-cost prior to guide token reduction. From a holistic spatio-temporal perspective, we further integrate question relevance from keyframe sampling to adaptively allocate token retention budgets. To facilitate evaluation, we construct EventBench, the first event-inclusive, human-annotated multimodal benchmark that covers diverse real-world scenarios. Beyond physical event cameras, EventSTU also supports general video understanding through simulated events. Comprehensive experiments show that EventSTU achieves 2.87× FLOPs reduction and 3.10× prefilling speedup over the strongest baseline while still improving performance.

Keywords: Event cameras · Efficient video understanding · Large language models

1 Introduction

Video Large Language Models (Video-LLMs) have made significant progress in complex video understanding tasks [21,23,25,33,43,60,61]. However, the massive number of frames in long videos produces tens of thousands of visual tokens. As the attention mechanism scales quadratically with token length, processing such sequences leads to prohibitive training and inference costs. In extreme cases, the input sequence may even exceed the maximum token capacity of LLMs, causing failure [39, 41, 59]. On the other hand, the redundant or irrelevant segments in long videos introduce substantial noise or uninformative tokens, which distract attention and hinder effective reasoning [29, 42]. These challenges raise a critical question: *how can we effectively reduce the number of visual tokens while maintaining model performance?*

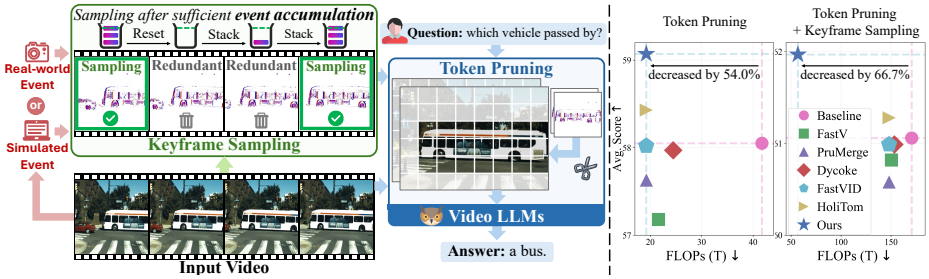

Fig. 1: Left: The key innovation of EventSTU lies in leveraging the change-triggered property of events to remove redundant frames, while utilizing their inherent visual saliency to prune less informative tokens. Notably, the framework can be driven by either real-world events or simulated events. **Right**: Our framework breaks the conventional efficiency-performance trade-off, reducing FLOPs by 66.7% yet still outperforming the original model. This holds for both pruning-only and combined settings.

In this work, we draw inspiration from biological vision systems, which achieve remarkable efficiency by prioritizing change and motion cues [1,5,12,16], enabling the system to focus neural resources on critical information and process massive visual information rapidly. Event cameras [6, 13, 30, 48, 50, 56] mimic this mechanism by asynchronously capturing only brightness changes at individual pixels, producing sparse yet information-rich event streams. Motivated by event-based vision, we propose EventSTU, a training-free framework for efficient long video understanding, as illustrated in Fig. 1, which leverages event guidance to eliminate temporal and spatial redundancy.

For temporal redundancy across frames, downsampling video frames is a straightforward way. However, the uniform sampling strategy adopted by most existing methods [21,43,61] always misses critical visual cues and causes key information loss. To address this limitation, we design a **C**oarse-to-**F**ine **S**ampling algorithm, termed C2FS. At the coarse stage, C2FS leverages the change-triggered nature of event cameras, where event density indicates information increment. Once sufficient events have accumulated across frames to signal a significant visual change, a keyframe is sampled. At the fine stage, C2FS further applies question-relevant sampling, selecting the most semantically relevant frames conditioned on the input question. The hierarchical sampling narrows the focus from redundancy elimination to semantic alignment, enabling efficient yet informative video understanding.

For spatial redundancy within frames, existing token pruning techniques can be broadly categorized into inner-LLM and outer-LLM methods. Inner-LLM pruning [8,10,36,37,42,57] operates after early LLM layers, yielding limited computational savings. Outer-LLM pruning [2,35,38,40,58] computes pairwise token similarities, resulting in quadratic overhead. To address these issues, we design a **Z**ero-cost **A**daptive **P**runing algorithm, termed ZAP, which mimics the human cognitive process from physical perception to semantic focus. Specifically, ZAP first exploits event sparsity as a natural saliency cue for physics-aware selection,

and then reuses attention scores to preserve non-salient yet semantically critical tokens for semantic-aware selection. Critically, rather than pruning each frame in isolation, ZAP integrates question relevance derived from C2FS to adaptively allocate per-frame retention budgets. This joint design enables a more informed and balanced token allocation with a holistic spatio-temporal view.

To evaluate event-guided video understanding, we construct EventBench, the first event-inclusive and human-annotated multimodal benchmark. EventBench provides high-quality event streams and RGB videos across diverse real-world scenarios, including 8 task types with different focuses. Multiple expert annotators carefully design 500 multiple-choice question-answer pairs that probe temporal localization, spatial grounding, and holistic video understanding.

More importantly, EventSTU also generalizes beyond real-world events to conventional videos through simulated event guidance. Given that existing Video-LLMs [21, 43, 61] typically process low-frame-rate inputs, fine-grained microsecond event details are unnecessary. Thus, we utilize a lightweight event simulation technique [31] that directly generates event frames [49] from adjacent image frames, offering on-the-fly simulated event guidance.

Our main contributions can be summarized as follows:

- We propose EventSTU, a training-free, event-guided framework for efficient spatio-temporal understanding, which comprises two elaborate strategies: C2FS to retrieve non-redundant, question-relevant keyframes, and ZAP to distill visually salient and semantically critical tokens.
- We construct EventBench, the first human-annotated multimodal benchmark with paired real event streams and RGB videos for evaluating event-guided video understanding.
- We extend EventSTU to conventional video understanding with simulated events, generalizing the event-inspired efficiency beyond real event data.
- Experimental results on four benchmarks demonstrate that EventSTU achieves state-of-the-art performance and efficiency under both real and simulated event settings. Consistent gains on five mainstream Video-LLMs further validate the strong generalizability of EventSTU.

2 Related Work

2.1 Token Reduction in Video-LLMs

Keyframe Sampling. The performance of Video-LLMs depends on the quality of the sampled frames. VideoAgent [46] and VideoTree [47] generate frame captions and utilize an LLM to select keyframes, while [17] trains a multimodal LLM for selection. BOLT [29] employs vision-language models to select question-relevant frames, and AKS [41] balances relevance with diversity. However, they rely on computationally expensive pre-processing. Instead, we leverage event density to efficiently eliminate redundant frames, reducing the substantial computational cost and achieving a superior efficiency-performance trade-off.

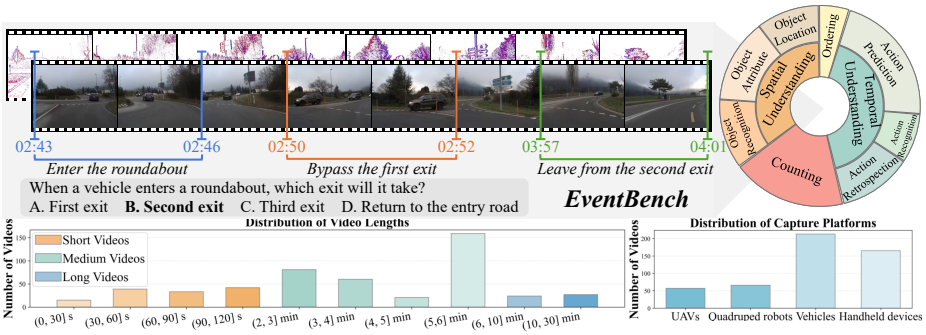

Fig. 2: Top: A sunburst and example of the proposed EventBench. **Bottom**: Statistics of EventBench including the distributions of video lengths and capture platforms.

Token Pruning. Visual token pruning has attracted increasing attention. FastV [8] prunes tokens via attention scores after early LLM layers, while PACT [10] further bypasses attention scores to enable FlashAttention [9] compatibility. DyCoKe [42] prunes key-value caches during decoding. However, these methods leave early layers unpruned, limiting FLOPs reduction. LLaVA-PruMerge [35] and VisionZip [58] merge tokens before the LLM based on token similarities, incurring quadratic computational overhead. In contrast, we leverage zero-cost event saliency to prune tokens before the LLM.

2.2 Event-based Understanding

Unlike mature image understanding, the event community is still in the early stages of developing universal models. EventCLIP [52] is the first to adapt CLIP for zero-shot event recognition. CEIA [55] and EventBind [62] leverage images to learn unified representations, addressing event-text data scarcity. To handle more complex interactive tasks, EventGPT [28] and EventVL [24] develop event-based LLMs, but are limited to short event streams (<100ms). These works aim to exploit the robustness of events for understanding under challenging conditions, while our work focuses on the efficiency of events.

3 EventBench Construction

EventBench is a multimodal benchmark featuring real-world event streams for evaluating Video-LLMs on event-guided long video understanding. The statistics of EventBench are shown in Fig. 2. Compared with concurrent benchmarks [27, 32] that contain only event streams and focus on event-only understanding, our benchmark provides paired event streams and RGB videos to support event-guided video understanding.

Data Collection. We curate a diverse collection of high-resolution event streams [7, 15, 20, 26], each paired with spatially aligned and temporally synchronized

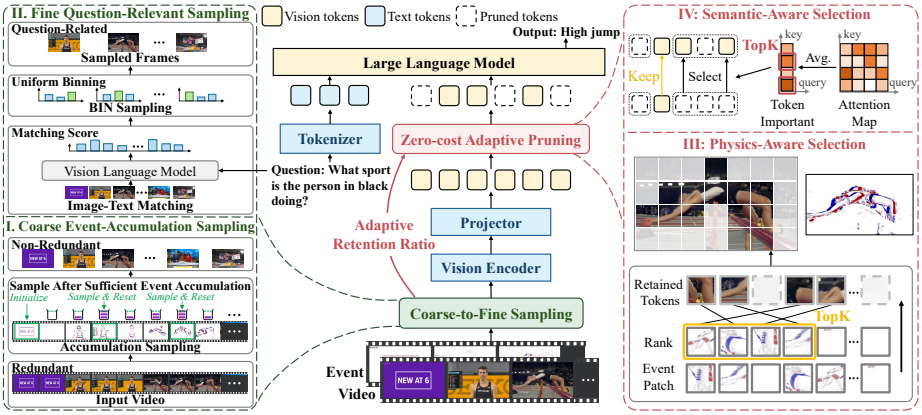

Fig. 3: Overview of EventSTU. It processes long videos through a multi-stage pipeline. First, coarse-to-fine sampling efficiently filters redundant frames and retrieves question-relevant keyframes. Subsequently, zero-cost adaptive pruning retains tokens that are either event-salient or semantically critical. This entire process culminates in a compact yet semantically dense visual representation, tailored for LLM inference.

RGB videos. To cover the application scenarios of event cameras, our benchmark spans four capture platforms: UAVs, ground vehicles, quadruped robots, and handheld devices, across three environments: indoor, outdoor daytime, and outdoor nighttime. Furthermore, after filtering short sequences, the remaining event streams average 4.3 minutes, with the longest reaching 23.5 minutes. This makes EventBench a valuable benchmark for long video understanding.

Question Design. We follow two fundamental design principles from existing multimodal benchmarks [11,19,44,51,53,54,63]: human annotation and multiple-choice format. These ensure evaluation reliability and quantifiability. Specifically, building on the collected data, multiple annotators meticulously design and rigorously validate 500 multiple-choice questions, covering 8 diverse tasks: object recognition, object attribute recognition, object localization, action recognition, action prediction, action retrospection, counting, and ordering. Among these, counting and ordering assess holistic video understanding, while the remaining tasks assess the localization and contextualization of specific segments. We believe that EventBench will advance the research of event-guided Video-LLMs by providing a comprehensive and reliable evaluation.

4 Event-Guided Video Understanding

4.1 Overview

To address the challenges in Video-LLMs, we propose EventSTU, a training-free framework that achieves efficient video understanding by coupling temporal keyframe sampling and spatial token pruning under a unified event-guided paradigm. An overview is illustrated in Fig. 3.

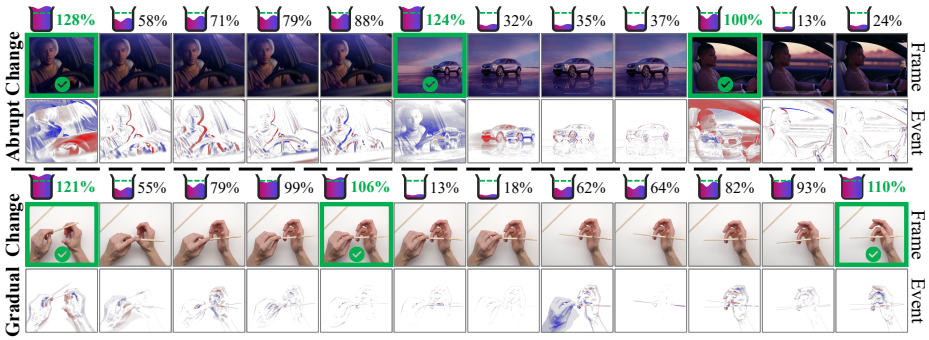

Fig. 4: Visualization of event-accumulation sampling in C2FS. After each sampling step, the accumulated event value is reduced by the threshold (100%). The two examples illustrate that redundant frames are effectively filtered out.

Given an RGB video sequence $\{F_t\}_{t=1}^{M}$, we additionally incorporate a sequence of event frames $\{E_t\}_{t=1}^{M-1}$, where each E_t encodes the accumulated brightness changes between consecutive frames F_{t-1} and F_t. These event frames can either originate from real sensors or be synthesized via simulators.

EventSTU comprises two components: coarse-to-fine sampling (C2FS) and zero-cost adaptive pruning (ZAP). In the first step, C2FS reduces the video from M original frames to M_c non-redundant frames, and then selects the M_f most question-relevant keyframes, where $M_f \ll M_c \ll M$. After tokenization and projection of the selected keyframes, ZAP sequentially applies physics-aware and semantic-aware selection with adaptive retention ratios to retain more important tokens. This yields compact yet semantically rich visual tokens.

Finally, the visual tokens and textual tokens derived from the question are concatenated and fed into the LLM. Through the synergy of C2FS and ZAP, EventSTU delivers both computational efficiency and strong video understanding performance, all without training or fine-tuning.

4.2 Coarse-to-fine Sampling

In long video understanding tasks, an effective keyframe sampling strategy must address two critical challenges: minimizing temporal redundancy and ensuring question relevance. Building on these principles, we design a coarse-to-fine sampling (C2FS) algorithm that progressively refines sampling from temporal redundancy removal to question relevance retrieval.

Coarse Event-Accumulation Sampling Stage. Event data is triggered by brightness changes, making it an ideal proxy for information increment. Leveraging this property, we propose an event-accumulation sampling algorithm.

Specifically, we first compute the event density sequence $\{e_t\}_{t=1}^{M-1}$, where e_t denotes the event count between consecutive frames F_{t-1} and F_t. By accumulating the event density across frames, we can measure the information increment

of the current frame relative to the last sampled keyframe. Once the accumulated value exceeds a threshold τ, the frame is sampled as a new keyframe, while intermediate frames are discarded as redundant. To handle videos with varying motion intensity, we adaptively compute the threshold τ for each video as

$$\tau = \frac{\sum_{t=1}^{M-1} e_t}{\lceil (M-1)S \rceil}, \tag{1}$$

where S denotes the predefined sampling ratio. The above process repeats until the entire sequence has been processed, producing a compact set of keyframes that effectively reduces temporal redundancy. As shown in Fig. 4, our strategy effectively handles both videos with abrupt changes and those with gradual changes. The detailed algorithm is provided in the supplementary material.

Fine Question-Relevant Sampling Stage. To further refine keyframe sampling, we incorporate question relevance through text-image similarity. In this stage, we first encode the question and the coarse-sampled frames using a text encoder and a visual encoder (e.g., BLIP [22] or CLIP [34]), respectively. Then, the cosine similarity scores are computed between their embeddings.

Instead of directly selecting M_f frames with top scores, we introduce a bin sampling strategy to ensure the selection spans a broad temporal range. Specifically, we uniformly divide the coarse-sampled frames into B bins based on their temporal position and select the frame with the highest similarity score from each bin. This method ensures that the selected frames maintain temporal diversity while simultaneously being highly relevant to the user question.

Putting things together, C2FS decouples keyframe sampling into redundancy removal and relevance retrieval. The coarse event-accumulation sampling serves as an efficient, content-agnostic filter, removing redundant frames and reducing the computational burden for subsequent fine sampling. The fine question-relevant sampling ensures the selected frames are highly relevant to the question.

4.3 Zero-Cost Adaptive Pruning

Even after reducing temporal redundancy across the keyframes, spatial redundancy within the selected frames remains a significant challenge. To address this limitation, we design a Zero-Cost Adaptive Pruning (ZAP) algorithm. ZAP operates in two phases: physics-aware selection and semantic-aware selection. These phases are applied sequentially to the set of visual tokens derived from the selected keyframes. Besides, ZAP adaptively allocates the retention ratio for each frame by integrating the question relevance from C2FS, thereby achieving holistic token pruning.

Physics-Aware Selection. Event data inherently provides more compact representations due to its sparse encoding of brightness changes. As shown in Fig. 5, this sparsity manifests in two ways: it filters out uninformative regions, such as the sky and roads, which trigger few events but occupy large portions of visual tokens. At the same time, it highlights changing regions that trigger more events and typically attract greater user interest. Building upon these insights, we introduce event-based physics-aware selection.

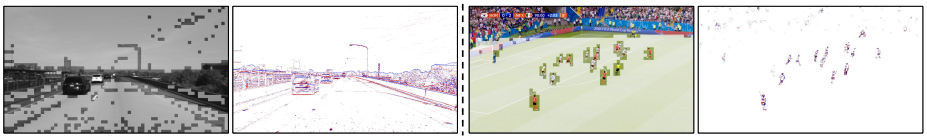

Fig. 5: Visualization of physics-aware selection in ZAP. Sparse events effectively filter out uninformative regions (e.g., sky and road) while highlighting changing regions.

Specifically, we partition an event frame into non-overlapping patches, each corresponding to an RGB visual token. We then compute the event density within each patch, using it as an importance score. For each RGB frame, we retain the top $R_p\%$ of tokens with the highest scores, effectively filtering out the uninformative regions. Our physics-aware selection is simple yet highly effective, introducing only a negligible 0.00027% increase in total computational cost.

Semantic-Aware Selection. Despite being effective, physics-aware selection may overlook non-salient yet semantically critical cues. To address this, we introduce semantic-aware selection by directly reusing the attention matrix from the last layer of the vision encoder to assess token importance. By averaging attention scores across the query dimension, we obtain a per-token importance measure. Visual tokens with higher average attention are deemed more semantically critical, while tokens with lower scores are less critical.

We retain the top $R_s\%$ of tokens with the highest attention scores, from the tokens not retained by physics-aware selection. Our semantic-aware selection benefits from the efficient reuse of the attention matrix, introducing only a negligible 0.000035% increase in total computational cost.

Adaptive Retention Ratio. To achieve holistic token pruning across keyframes, we introduce a strategy that adaptively allocates retention ratios based on their relevance to the input question. More relevant keyframes retain more tokens, while less relevant keyframes undergo more aggressive pruning.

Mathematically, the total allocable token budgets $N_{\text{alloc}} = M_f \cdot N \cdot (R - B)$, where M_f is the number of keyframes, N is the number of tokens per frame, R is the expected overall retention ratio, and B is the base retention ratio (empirically set to 5% to avoid over-pruning, with $B \leq R$). The adaptive retention ratio R^t for frame F_t is defined as:

$$R^t = \min\left(\frac{N_{\text{alloc}} \cdot s^t + B \cdot N}{N}, 1\right), \tag{2}$$

where s^t is the similarity score to the question, derived from C2FS and normalized via softmax. In practice, if the computed R^t exceeds 1, all tokens of frame F_t are retained, and any excess token budget is returned to the total allocation.

To balance the expected physics-aware retention ratio R_p and semantic-aware retention ratio R_s, we set:

$$(R_p^t, R_s^t) = \begin{cases} (R^t, 0) & \text{if } R^t \leq R_p, \\ (R_p, \frac{R^t - R_p}{1 - R_p}) & \text{otherwise,} \end{cases} \tag{3}$$

where R_p^t and R_s^t denote the adaptive retention ratios in physics-aware and semantic-aware selection for frame F_t, respectively.

5 Extension to General Video Understanding

Event data is known for its superior efficiency, prompting the question: *Can our framework be driven by simulated events instead of relying on physical event cameras?* Traditional event simulation techniques [14, 18, 64] typically interpolate video to extremely high frame rates, introducing unacceptable delays of up to several minutes. To avoid this, we utilize a lightweight event simulation method [31] tailored to general video understanding, leveraging frame differences to simulate event frames without complex interpolation.

We simulate events by calculating intensity differences between consecutive RGB frames, F_t and F_{t-1}. After applying reverse gamma correction γ^{-1}, we compute the intensity change $\triangle L_{t,t-1}$ as:

$$\triangle L_{t,t-1} = \log(\gamma^{-1}(F_t)) - \log(\gamma^{-1}(F_{t-1})). \tag{4}$$

Next, we calculate the number of positive and negative events at each pixel using predefined thresholds C_p and C_n:

$$N_p = \left\lfloor \frac{[\triangle L_{t,t-1}]_+}{C_p} \right\rfloor, N_n = \left\lfloor \frac{[- \triangle L_{t,t-1}]_+}{C_n} \right\rfloor, \tag{5}$$

where $[\cdot]_+$ denotes $max(0, \cdot)$. These events are then used to construct a 2D event frame with the total event count $N_p + N_n$ at each pixel. Importantly, this method eliminates the need for microsecond-level interpolation, which is computationally expensive and unnecessary, while still capturing the second-level changes needed for video understanding. Additionally, to overcome the limitations of physics-aware selection in completely static scenes (e.g., slide-style videos), we introduce a shake trick during simulation, as detailed in the supplementary material.

With this efficient simulation, EventSTU breaks the dependence on physical event sensors, enabling straightforward extension to general video understanding.

6 Experiment

6.1 Experimental Settings

Benchmarks. We evaluate our method on EventBench with real event data and three widely-used general video benchmarks with simulated event data: LongVideoBench [51], Video-MME [11], and MLVU [63]. These benchmarks provide a comprehensive testbed for assessing the effectiveness and generalization of our method with and without physical event cameras.

Comparison Methods. Our method is compared in three settings: keyframe sampling, token pruning, and their combination. For keyframe sampling, we compare our C2SF against two latest training-free methods, AKS [41] and BOLT [29].

Table 1: Comparison on LLaVA-OneVision [21] across four benchmarks: EventBench (EB) with real event data, and Video-MME (V-MME), LongVideoBench (LVB), and MLVU with simulated event data.

Method	Retention Ratio	FLOPs (T)↓	FLOPs Ratio↓	EB↑	V-MME↑	LVB↑	MLVU↑	Avg.↑
LLaVA-OV-7B [21]	100%	41.71	100%	52.50	58.48	56.40	64.81	58.05
Keyframe Sampling								
BOLT [29]	100%	170.6	409%	-	59.90	59.60	66.80	-
AKS [41]	100%	170.6	409%	54.09	60.44	60.28	69.41	61.06
C2FS (Ours)	100%	**79.34**	**190%**	**55.09**	**60.96**	**60.43**	**69.87**	**61.59**
Token Pruning								
FastV [8]	50%	21.61	51.8%	50.90	57.74	56.10	63.98	57.18
PruMerge [35]	50%	19.24	46.1%	51.90	58.04	55.95	64.63	57.63
DyCoke [42]	50%	24.45	58.6%	52.50	59.04	55.80	64.54	57.97
FastVID [38]	50%	19.21	46.1%	52.69	58.33	56.55	64.49	58.02
HoliTom [37]	50%	19.21	46.1%	52.69	58.52	56.62	**65.87**	58.43
ZAP (Ours)	50%	**19.20**	**46.0%**	**54.09**	**59.15**	**57.37**	65.69	**59.07**
Keyframe Sampling + Token Pruning								
AKS [41] + FastV [8]	50%	150.5	361%	53.49	60.22	60.28	69.27	60.82
AKS [41] + PruMerge [35]	50%	148.1	355%	52.10	60.48	59.84	69.92	60.58
AKS [41] + DyCoke [42]	50%	153.3	368%	53.49	60.56	59.91	70.01	60.99
AKS [41] + FastVID [38]	50%	148.1	355%	53.89	60.44	59.99	69.64	60.99
AKS [41] + HoliTom [37]	50%	148.1	355%	54.29	**60.78**	59.99	70.06	61.28
EventSTU (Ours)	50%	**56.83**	**136%**	**55.09**	60.70	**61.11**	**70.98**	**61.97**
AKS [41] + FastV [8]	30%	143.4	344%	51.10	58.89	58.56	69.41	59.49
AKS [41] + PruMerge [35]	30%	140.2	336%	52.89	59.89	57.59	69.69	60.02
AKS [41] + DyCoke [42]	30%	147.1	353%	53.89	60.41	59.16	70.61	61.02
AKS [41] + FastVID [38]	30%	140.2	336%	52.30	60.67	60.14	69.87	60.74
AKS [41] + HoliTom [37]	30%	140.2	336%	51.90	60.78	59.01	70.84	60.63
EventSTU (Ours)	30%	**48.93**	**117%**	**54.49**	**61.07**	**60.73**	**71.16**	**61.86**

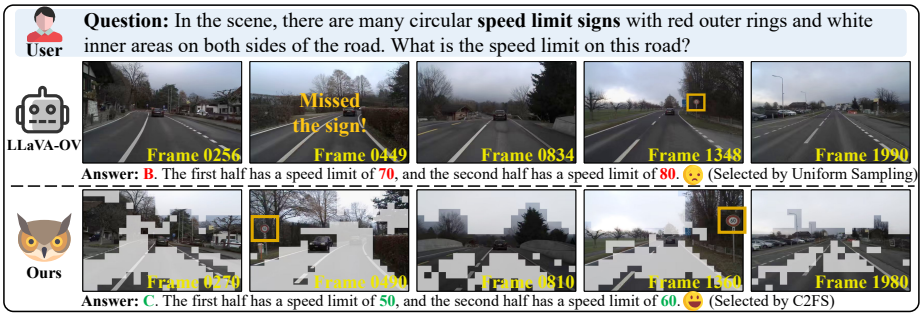

Fig. 6: Visualization on LLaVA-OV. The original model misses a keyframe, while our method captures sufficient keyframes and prunes uninformative regions.

Since BOLT is not publicly available, we compare against its reported results. For token pruning, we compare our ZAP against five strong training-free methods: FastV [8], LLaVA-PruMerge [35], DyCoKe [42], FastVID [38], and HoliTom [37]. For the combined setting, we compare our complete framework against AKS [41] paired with each token pruning method.

Implementation Details. We implement our methods on LLaVA-OneVision-7B [21] and Qwen3-VL [3] using 8 NVIDIA RTX 3090 GPUs. For LLaVA-

Table 2: Comparison on Qwen3-VL [3] across four benchmarks: EventBench (EB) with real event data, and Video-MME (V-MME), LongVideoBench (LVB), and MLVU with simulated event data.

Method	Retention Ratio	FLOPs (T)↓	FLOPs Ratio↓	EB↑	V-MME↑	LVB↑	MLVU↑	Avg.↑
Qwen3-VL-8B [3]	100%	169.1	100%	69.06	67.52	61.33	68.49	66.60
Keyframe Sampling								
AKS [41]	100%	297.9	176%	69.46	68.48	**62.60**	76.45	69.25
C2FS (Ours)	100%	**206.7**	**122%**	**71.26**	**69.56**	**62.60**	**77.09**	**70.13**
Token Pruning								
FastV [8]	50%	75.30	44.5%	67.07	66.85	**59.84**	67.99	65.43
PruMerge [35]	50%	66.79	**39.5%**	66.67	65.93	58.86	68.22	64.92
FastVID [38]	50%	66.83	**39.5%**	67.47	67.00	59.76	67.99	65.55
ZAP (Ours)	50%	**66.77**	**39.5%**	**68.46**	**67.41**	**59.84**	**68.40**	**66.03**
Keyframe Sampling + Token Pruning								
AKS [41] + FastV [8]	50%	204.2	121%	67.47	67.81	62.45	76.40	68.53
AKS [41] + PruMerge [35]	50%	195.7	116%	67.47	66.85	61.78	75.80	67.98
AKS [41] + FastVID [38]	50%	195.7	116%	69.26	68.70	**63.05**	76.17	69.30
EventSTU (Ours)	50%	**104.4**	**62%**	**70.46**	**69.30**	62.83	**76.49**	**69.77**
AKS [41] + FastV [8]	30%	176.0	104%	64.87	67.81	61.18	75.76	67.41
AKS [41] + PruMerge [35]	30%	164.9	98%	64.67	65.37	59.24	74.84	66.03
AKS [41] + FastVID [38]	30%	164.9	98%	66.67	66.89	62.08	75.76	67.85
EventSTU (Ours)	30%	**73.64**	**44%**	**70.06**	**68.78**	**62.15**	**76.17**	**69.29**

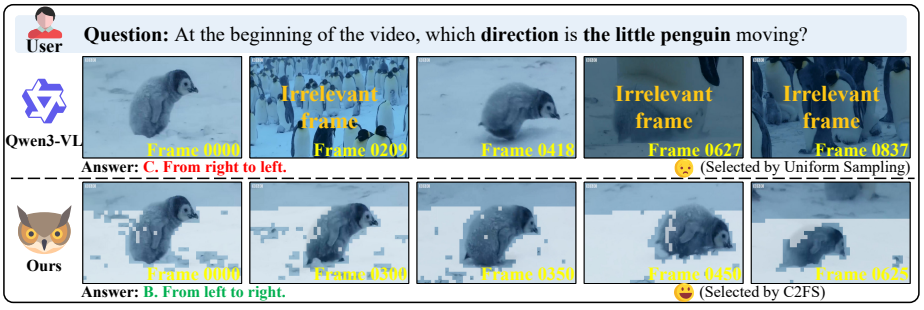

Fig. 7: Visualization on Qwen3-VL. The original model is affected by irrelevant frames, while ours captures continuous keyframes and prunes uninformative regions.

OneVision, we use 32 input frames (i.e., our fine-sampling target M_f) with $N = 196$ tokens per frame. For Qwen3-VL, we also use 32 input frames with a maximum resolution of $640 \times 28 \times 28$, which dynamically determines N per frame. We set the coarse frame sampling ratio S to 25%. The overall retention ratio R is treated as the main experimental variable. Detailed hyperparameter analysis is provided in the supplementary material.

6.2 Comparison with State-of-the-Art Methods

As shown in Tab. 1, our EventSTU achieves state-of-the-art performance and efficiency across four benchmarks and three settings on LLaVA-OneVision [21].

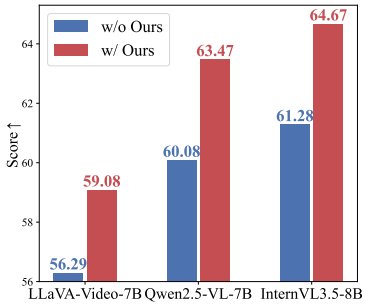

Fig. 8: Generalization to other Video-LLMs with 50% retention ratio on EventBench.

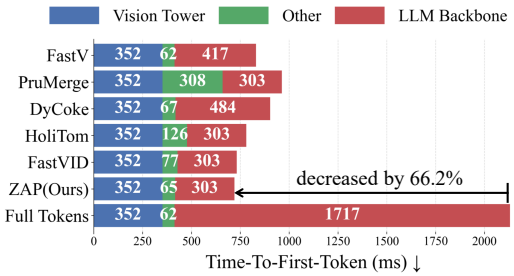

Fig. 9: Latency Analysis. "Other" indicates token pre-processing time. Our ZAP achieves the lowest Time-to-First-Token.

Keyframe Sampling. BOLT [29] and AKS [41] improve performance but incur a substantial 309% increase in FLOPs. In contrast, our C2FS eliminates temporal redundancy, reducing the number of processed frames. Compared to AKS, C2FS requires only 46.5% of its FLOPs while still surpassing it by 0.53 points in average performance, making keyframe sampling more practical.

Token Pruning. At the same retention ratio, ZAP achieves the highest average performance, the lowest FLOPs, and the minimal inference latency (refer to Sec. 6.4). For example, ZAP surpasses DyCoke [42] by 1.10 points in average performance while requiring only 78.5% of its FLOPs.

Keyframe Sampling + Token Pruning. By integrating C2FS and ZAP, EventSTU facilitates holistic spatio-temporal token allocation. This synergistic design yields impressive results: at a 30% retention ratio, EventSTU surpasses the original model by a significant 3.81 points in average performance.

Fig. 6 further illustrates this advantage: the original model misses a keyframe and is distracted by irrelevant sky and road regions, while EventSTU captures sufficient keyframes and prunes distracting areas, highlighting the queried speed limit signs. These results collectively demonstrate the effectiveness of EventSTU in leveraging the spatio-temporal efficiency of event data.

6.3 Generalization Across Different Video-LLMs

As shown in Tab. 2, EventSTU also achieves state-of-the-art performance and efficiency on Qwen3-VL [3]. Due to the naive dynamic resolution strategy [43], Qwen3-VL consumes more visual tokens than LLaVA-OneVision [21], thereby intensifying the demand for token reduction. Meeting this demand, EventSTU demonstrates exceptional efficacy on Qwen3-VL, breaking the conventional efficiency–performance trade-off. At a 30% retention ratio, EventSTU improves average performance by 2.69 points over the original model while requiring only 44% of its FLOPs, notably without any training or fine-tuning.

Moreover, Fig. 8 shows that EventSTU further generalizes effectively to three additional Video-LLMs, including LLaVA-Video [61], Qwen2.5-VL [4], and

Table 3: Comparison with alternative sampling strategies. Please refer to Sec. 6.5 for abbreviation explanations.

Method		EventBench↑	LongVideoBench↑
Coarse Sampling	UNI	53.09	59.54
	TOP	54.09	58.94
	ACC	**55.09**	**60.43**
Fine Sampling	TOP	53.89	60.21
	BIN	**55.09**	**60.43**

Table 4: Ablation study on ZAP. "Sem." denotes semantic-aware selection, "Phy." denotes physics-aware selection, and "Ada." denotes adaptive retention ratio.

Sem.	Phy.	Ada.	EventBench↑	LongVideoBench↑
✓			52.10	58.12
	✓		53.29	58.64
✓	✓		54.09	59.54
✓	✓	✓	**55.09**	**61.11**

Intern3.5-VL [45]. Across these diverse architectures, EventSTU achieves an average improvement of 3.19 points, demonstrating strong cross-model generalization and broad applicability.

6.4 Latency Analysis

Fig. 9 presents a latency comparison between our ZAP and existing token-pruning methods. Since our focus is on the pre-filling stage, we measure the time-to-first-token (TTFT). Specifically, FastV [8] and DyCoKe [42] fail to reduce latency in the early LLM layers due to their design. Meanwhile, LLaVA-PruMerge [35] and HoliTom [37] rely on computationally intensive pre-processing, such as similarity calculation and clustering, significantly increasing the "other" time.

In contrast, our ZAP avoids these limitations by leveraging zero-cost guidance from event data and reused attention, introducing only 1.24 ms of additional time. Due to the lightweight design of the frame-difference-based simulator, event simulation adds only 1.52 ms. Overall, ZAP reduces TTFT by 66.2% compared to the original model — the largest reduction among all compared methods.

6.5 Ablation Study

Comparison with Alternative Sampling Strategies. As shown in Tab. 3, we compare alternative strategies for coarse and fine sampling on EventBench and LongVideoBench. At the coarse sampling stage, "UNI" (uniform sampling) performs worst because its sampling decisions are independent of information density, leading to under-sampling during periods with rapid change. "TOP" (selecting frames with the highest event density) also struggles, as it focuses solely on instantaneous density and completely overlooks low-density regions that may contain important information. In contrast, "ACC" (accumulation sampling) adaptively selects frames based on accumulated event density, thereby capturing both abrupt and gradual changes and achieving the best performance.

At the fine sampling stage, "TOP" (sampling frames with the highest image-text similarity) spans a narrow temporal window. Despite enabling localization of the queried segments, it lacks sufficient context to answer the question. In contrast, "BIN" (bin sampling) effectively balances question relevance and temporal coverage, significantly outperforming "TOP" by 1.20 points on EventBench.

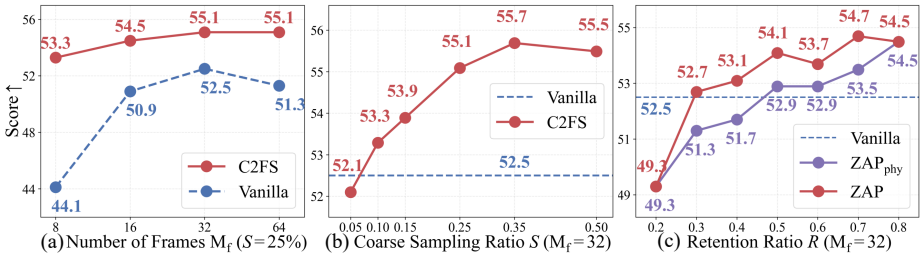

Fig. 10: Analysis of number of frames (a), coarse sampling ratio (b) in C2FS, and retention ratio (c) in ZAP. "Vanilla" denotes the original model. "ZAP$_{\text{phy}}$" denotes using physics-aware selection only.

Ablation Study on ZAP. As shown in Tab. 4, combining physics-aware and semantic-aware selection outperforms using either method alone, demonstrating the importance of retaining both visually salient and semantically critical tokens simultaneously. Additionally, we observe that adaptive retention ratio improves average performance by an impressive 1.28 points, illustrating the effectiveness of optimizing token budgets from a holistic spatio-temporal perspective.

Analysis on Number of Frames. Fig. 10 (a) demonstrates the consistent superiority of C2FS over the original model. The improvement is particularly evident with a few frames ($M_f = 8$), confirming that C2FS effectively eliminates redundant and irrelevant frames, thereby enhancing the signal-to-noise ratio.

Analysis on Coarse Sampling Ratio. S controls the number of coarse sampled frames, determining the degree of redundancy removal. As shown in Fig. 10 (b), the highest coarse sampling rate ($S = 0.50$) fails to yield the best results, despite providing more candidates for fine sampling. This indicates that redundancy removal narrows the search scope, enhancing question relevant retrieval.

Analysis on Retention Ratio. Fig. 10 (c) shows that full ZAP consistently outperforms ZAP$_{\text{phy}}$ (using physics-aware selection only). Interestingly, ZAP with high retention ratio ($R = 0.7$) significantly outperforms the original model by 2.19 points. This suggests that proper pruning can filter out noisy signals and enhance the understanding capability of LLMs.

7 Conclusion

This work proposes a training-free, event-guided framework to address the challenges of long video understanding in Video-LLMs. It is built on a dual exploitation of event data: a coarse-to-fine sampler that leverages the change-triggered property of events to reduce temporal redundancy, and an adaptive token pruner that utilizes their visual saliency to filter uninformative tokens. We validate our method on EventBench, our newly constructed benchmark featuring real events, as well as on three general video benchmarks with simulated events. The comprehensive empirical evaluation across four benchmarks confirms the effectiveness of our event-guided strategy, paving the way for event-guided LLMs.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 62472399, and the Fundamental Research Funds for the Central Universities under Grant WK2100000059.

References

1. Albright, T.D., Stoner, G.R.: Visual motion perception. Proceedings of the National Academy of Sciences **92**(7), 2433–2440 (1995)
2. Alvar, S.R., Singh, G., Akbari, M., Zhang, Y.: Divprune: Diversity-based visual token pruning for large multimodal models. In: CVPR (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
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. Berry, M.J., Brivanlou, I.H., Jordan, T.A., Meister, M.: Anticipation of moving stimuli by the retina. Nature **398**(6725), 334–338 (1999)
6. Chakravarthi, B., Verma, A.A., Daniilidis, K., Fermuller, C., Yang, Y.: Recent event camera innovations: A survey. In: ECCV (2024)
7. Chaney, K., Cladera, F., Wang, Z., Bisulco, A., Hsieh, M.A., Korpela, C., Kumar, V., Taylor, C.J., Daniilidis, K.: M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In: CVPR workshop (2023)
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: ECCV (2024)
9. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. In: NeurIPS (2022)
10. Dhouib, M., Buscaldi, D., Vanier, S., Shabou, A.: Pact: Pruning and clustering-based token reduction for faster visual language models. In: CVPR (2025)
11. 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: CVPR (2025)
12. Gabbiani, F., Krapp, H.G., Laurent, G.: Computation of object approach by a wide-field, motion-sensitive neuron. Journal of Neuroscience **19**(3), 1122–1141 (1999)
13. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(1), 154–180 (2020)
14. Gehrig, D., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Video to events: Recycling video datasets for event cameras. In: CVPR (2020)
15. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3), 4947–4954 (2021)
16. Gollisch, T., Meister, M.: Eye smarter than scientists believed: neural computations in circuits of the retina. Neuron **65**(2), 150–164 (2010)

17. 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: CVPR (2025)
18. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: CVPR workshop (2021)
19. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
20. Kim, T., Jeong, J., Cho, H., Jeong, Y., Yoon, K.J.: Towards real-world event-guided low-light video enhancement and deblurring. In: ECCV (2024)
21. 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)
22. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
23. 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)
24. Li, P., Lu, Y., Song, P., Li, W., Yao, H., Xiong, H.: Eventvl: Understand event streams via multimodal large language model. arXiv preprint arXiv:2501.13707 (2025)
25. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
26. Liu, H., Peng, S., Zhu, L., Chang, Y., Zhou, H., Yan, L.: Seeing motion at nighttime with an event camera. In: CVPR (2024)
27. Liu, S., Li, J., Zhao, G., Zhang, Y., Ji, X.: Eventbench: Towards comprehensive benchmarking of event-based mllms. arXiv preprint arXiv:2511.18448 (2025)
28. Liu, S., Li, J., Zhao, G., Zhang, Y., Meng, X., Yu, F.R., Ji, X., Li, M.: Eventgpt: Event stream understanding with multimodal large language models. In: CVPR (2025)
29. Liu, S., Zhao, C., Xu, T., Ghanem, B.: Bolt: Boost large vision-language model without training for long-form video understanding. In: CVPR (2025)
30. Liu, Y., Weng, W., Zhang, Y., Xiong, Z.: S2d-lfe: Sparse-to-dense light field event generation. In: CVPR (2025)
31. Lou, H., Liang, J., Teng, M., Wang, Y., Shi, B.: V2v: Scaling event-based vision through efficient video-to-voxel simulation. In: NeurIPS (2026)
32. Lou, H., Zhou, J., Zhang, Y., Li, B., Wang, Y., Ye, G., Shi, B.: Reconstruction as a bridge for event-based visual question answering. arXiv preprint arXiv:2512.11510 (2025)
33. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: ACL (2024)
34. 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: ICML (2021)
35. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In: ICCV (2025)
36. Shao, K., Keda, T., Zhang, K., Feng, S., Cai, M., Shang, Y., You, H., Qin, C., Sui, Y., Wang, H.: A survey of token compression for efficient multimodal large language models. Transactions on Machine Learning Research (2026)

37. 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)
38. 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)
39. 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: ICML (2025)
40. Sun, B., Zhao, J., Wei, X., Hou, Q.: Llava-scissor: Token compression with semantic connected components for video llms. arXiv preprint arXiv:2506.21862 (2025)
41. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: CVPR (2025)
42. Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: Dycoke: Dynamic compression of tokens for fast video large language models. In: CVPR (2025)
43. 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)
44. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: ICCV (2025)
45. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
46. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: ECCV (2024)
47. 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: CVPR (2025)
48. Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: ICCV (2021)
49. Weng, W., Zhang, Y., Xiong, Z.: Boosting event stream super-resolution with a recurrent neural network. In: ECCV (2022)
50. Weng, W., Zhang, Y., Xiong, Z.: Event-based blurry frame interpolation under blind exposure. In: CVPR (2023)
51. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. In: NeurIPS (2024)
52. Wu, Z., Liu, X., Gilitschenski, I.: Eventclip: Adapting clip for event-based object recognition. arXiv preprint arXiv:2306.06354 (2023)
53. Xiong, J., Pu, Y., Tang, J., Niu, Y.: Priorzero: Bridging language priors and world models for decision making. arXiv preprint arXiv:2605.12289 (2026)
54. Xiong, J., Wang, Y., Zhao, W., Liu, C., Yin, B., Zhou, W., Li, H.: Docr1: Evidence page-guided grpo for multi-page document understanding. In: AAAI (2026)
55. Xu, W., Weng, W., Zhang, Y., Xiong, Z.: Ceia: Clip-based event-image alignment for open-world event-based understanding. In: ECCV workshop (2024)
56. Xu, W., Weng, W., Zhang, Y., Xu, R., Xiong, Z.: Event-boosted deformable 3d gaussians for dynamic scene reconstruction. In: ICCV (2025)
57. Yang, C., Sui, Y., Xiao, J., Huang, L., Gong, Y., Li, C., Yan, J., Bai, Y., Sadayappan, P., Hu, X., et al.: Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In: CVPR (2025)

58. 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: CVPR (2025)
59. Yu, S., JIN, C., Wang, H., Chen, Z., Jin, S., ZUO, Z., XIAOLEI, X., Sun, Z., Zhang, B., Wu, J., et al.: Frame-voyager: Learning to query frames for video large language models. In: ICLR (2025)
60. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: EMNLP (2023)
61. 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)
62. Zhou, J., Zheng, X., Lyu, Y., Wang, L.: Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. In: ECCV (2024)
63. Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv e-prints pp. arXiv–2406 (2024)
64. Ziegler, A., Teigland, D., Tebbe, J., Gossard, T., Zell, A.: Real-time event simulation with frame-based cameras. In: ICRA (2023)