EVIQ: HIERARCHICAL QUERY-CONDITIONED EVIDENCE SELECTION FOR EFFICIENT VIDEOLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

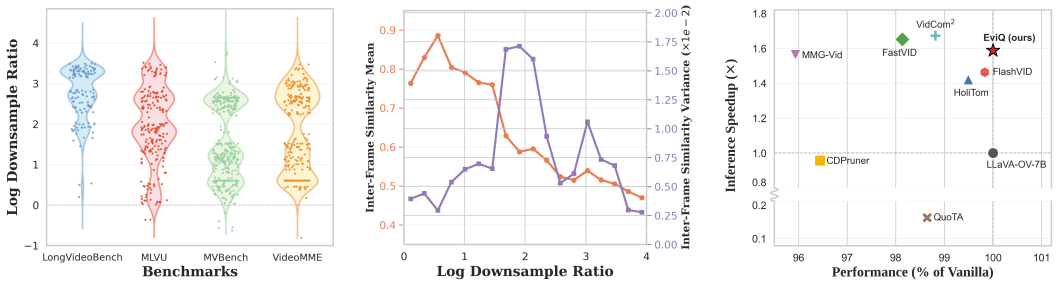
Video large language models (VideoLLMs) have achieved strong performance in video understanding, yet their large visual token budgets introduce substantial inference latency and memory consumption. Most existing training-free token compression methods rely primarily on visual cues, potentially overlooking query-critical evidence scattered across sparsely sampled frames, while some query-aware approaches incur additional computation through within-LLM attention or auxiliary models. Instead, we reveal that cosine similarity between projected visual tokens and embedded query tokens provides a lightweight relevance signal both across and within sparsely sampled frames. Building on this observation, we propose **EviQ**, a training-free pre-LLM token pruning framework that constructs a compact query-conditioned evidence set in two stages. **EviQ** first routes the global token budget across frames according to frame-level query relevance, and then selects tokens within each frame by jointly preserving informative visual context and promoting complementary coverage of query semantics. Extensive experiments on diverse VideoLLMs and multiple benchmarks demonstrate that **EviQ** achieves state-of-the-art accuracy-efficiency trade-offs among training-free token compression methods. Notably, **EviQ** preserves **99.9%** of vanilla LLaVA-OV-7B accuracy using only 25% of visual tokens, while achieving a 1.59× speedup and reducing GPU memory consumption by 8.7%. Code is released.

1 INTRODUCTION

Video large language models (VideoLLMs) Li et al. (2024a); Zhang et al. (2024); Bai et al. (2025) have recently achieved remarkable progress in video comprehension, temporal reasoning, and video question answering by extending multimodal LLMs Liu et al. (2023; 2024); Wang et al. (2024); Chen et al. (2024) with temporal visual signals. To capture sufficient spatiotemporal context, these models typically encode densely sampled video frames into a large number of visual tokens, which are then projected into the LLM for reasoning and generation. Although effective, this paradigm introduces a substantially larger token budget than image-based multimodal models, leading to significant inference latency and memory consumption. Such efficiency bottlenecks have become a major obstacle to deploying VideoLLMs in practical long-form video understanding scenarios.

To alleviate this bottleneck, a growing body of research has explored training-free visual token compression. Existing methods can be broadly grouped into two categories. The first relies mainly on visual cues, such as inter-frame similarity Fu et al. (2025b); Shao et al. (2026), temporal variation Shen et al. (2026); Ma et al. (2026), and token-level uniqueness or diversity Liu et al. (2025); Fan et al. (2026). Although effective at removing generic visual redundancy, these methods are largely query-agnostic. In long videos, the uniform downsampling commonly adopted by VideoLLMs weakens temporal continuity, as shown in Figure 1a and Figure 1b. Consequently, visually driven compression may spend the limited token budget on distinct but question-irrelevant background content, while pruning sparse evidence required by the query. The second category incorporates query information, but often at the cost of efficiency or deployability. Attention-based pruning inside the LLM Tao et al. (2025); Huang et al. (2025) yields limited speedups and requires adaptation to efficient attention operators such as FlashAttention Dao et al. (2022); Dao (2024), while auxiliary-network-based relevance estimation Luo et al. (2026) introduces extra computation and model-capacity dependence. Recent methods designed for image understanding also exploit query relevance for pre-LLM token selection Zhang et al. (2026); Li et al. (2026). However, these meth-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b) (c)

Figure 1: Motivations and highlights of **EviQ**. (a) Log downsampling ratio of video benchmarks, defined as $\log_{10}(\text{frames}/32)$. (b) Mean and variance of inter-frame cosine similarity across different downsampling ratios. Mean similarity exhibits an overall downward trend as the downsampling ratio increases, suggesting weaker visual continuity between sampled frames under more aggressive downsampling. (c) Accuracy-efficiency trade-off on LLaVA-OV-7B at a 25% retention ratio. **EviQ** preserves **99.9%** of the full-model accuracy while achieving a 1.59× inference speedup, delivering state-of-the-art performance among training-free token compression methods.

ods do not explicitly address how to distribute token budgets across sparsely sampled video frames while preserving complementary evidence within each frame. These limitations call for a pre-LLM framework that uses lightweight query guidance for both frame-level budget allocation and intra-frame selection, preserving compact query-conditioned visual evidence.

To address these challenges, we propose **EviQ**, a training-free pre-LLM hierarchical token pruning framework for efficient VideoLLMs. Building on prior studies Zhang et al. (2026); Li et al. (2026), we reveal that cosine similarity between projected visual tokens and embedded query tokens in the shared embedding space captures meaningful query relevance across and within sparsely sampled frames, motivating its use as a lightweight signal for both frame-level budget allocation and intra-frame selection in our method. **Frame-Level Query-Conditioned Budget Routing** aggregates token-query similarities into frame relevance scores and converts them into a soft budget distribution under the global retention constraint, assigning more capacity to frames likely to contain answer-related evidence. Given each frame budget, **Intra-Frame Query-Conditioned Evidence Selection** selects tokens as a subset by combining **Context-Preserving Evidence (CPE)**, which favors visually informative tokens as supporting scene context, and **Query-Complete Evidence (QCE)**, which promotes complementary coverage of query semantics. The intra-frame objective is monotone and submodular, enabling efficient greedy optimization with bounded approximation error.

Extensive experiments on multiple benchmarks Li et al. (2024b); Wu et al. (2024); Zhou et al. (2025); Fu et al. (2025a) and representative VideoLLMs Li et al. (2024a); Zhang et al. (2024); Bai et al. (2025) demonstrate that **EviQ** achieves state-of-the-art accuracy-efficiency trade-offs among training-free token compression methods as shown in Fig 1c. In particular, on LLaVA-OV-7B, **EviQ** retains **99.9%** of the full-model accuracy using only 25% visual tokens, while achieving a 1.59× speedup and an 8.7% GPU memory reduction. Moreover, since **EviQ** operates before LLM ingestion, it is orthogonal to FlashAttention Dao et al. (2022); Dao (2024) and generalizes to diverse video sampling strategies Tang et al. (2025); Sun et al. (2025).

Our contributions can be summarized as follows:

- We reveal that, in sparsely sampled videos, cosine similarity between projected visual tokens and embedded query tokens captures meaningful query relevance both across frames and within individual frames.
- We propose **EviQ**, a training-free query-conditioned token pruning framework that routes token budgets across frames and selects intra-frame evidence subsets by jointly preserving informative visual context and covering complementary query semantics, with an efficient greedy optimizer enjoying bounded approximation error.
- Extensive experiments across multiple VideoLLMs, benchmarks, retention ratios, and sampling strategies demonstrate that **EviQ** achieves state-of-the-art accuracy-efficiency trade-offs among training-free token compression methods.

2 Related Work

2.1 Video Large Language Models

Video large language models (VideoLLMs) extend multimodal LLMs to video understanding by incorporating temporal visual information. Representative models, such as LLaVA-OneVision Li et al. (2024a), LLaVA-Video Zhang et al. (2024), and Qwen3-VL Bai et al. (2025), have demonstrated strong performance in video comprehension, temporal reasoning, and video question answering. These models typically encode sampled video frames into dense visual tokens and feed them into the LLM together with textual queries. Compared with images, videos introduce substantially larger visual token budgets due to their temporal dimension, leading to high prefilling latency, memory consumption, and decoding overhead. This efficiency bottleneck has motivated recent studies on token pruning and compression for VideoLLMs.

2.2 Visual Token Pruning for VideoLLMs

To improve inference efficiency without additional training, recent studies have investigated training-free token pruning and compression for VideoLLMs. These methods can be broadly categorized according to whether the user's textual query is exploited during compression.

Vision-cue-based compression. The first line relies mainly on visual cues, including inter-frame similarity Fu et al. (2025b); Shao et al. (2026), temporal variation Shen et al. (2026); Ma et al. (2026); Ju et al. (2026), and token-level uniqueness or diversity Liu et al. (2025); Fan et al. (2026). At the frame level, these methods often depend on temporal continuity among sampled frames, which can be weakened by the aggressive downsampling used in modern VideoLLMs Tang et al. (2025); Zhang et al. (2025b); Sun et al. (2025); Zhu et al. (2025); Chen et al. (2026). As shown in Figure 1a and Figure 1b, long-video benchmarks such as LongVideoBench Wu et al. (2024) and MLVU Zhou et al. (2025) require substantial temporal downsampling. Larger downsampling ratios are associated with lower average cosine similarity between adjacent sampled frames, indicating weaker local temporal continuity in sparsely sampled inputs. This observation suggests that compression strategies relying on redundancy between neighboring frames may become less reliable under aggressive temporal downsampling. At the intra-frame level, diversity-driven selection tends to produce a broad but unfocused visual summary. For video question answering tasks (VideoQA), where answer evidence is often sparse and localized, the limited token budget may be spent on visually distinct but query-irrelevant background variations while query-critical tokens are pruned.

Query-aware compression. The second line incorporates textual queries into the compression process. One branch uses shallow-layer attention scores between query and visual tokens within the LLM to prune less relevant visual tokens in subsequent layers Tao et al. (2025); Huang et al. (2025). However, obtaining relevance signals after LLM computation has begun limits acceleration and requires non-trivial adaptations to efficient attention operators such as FlashAttention Dao et al. (2022); Dao (2024). Another branch estimates query relevance using an external auxiliary model Luo et al. (2026), introducing additional computational overhead and making compression quality dependent on the auxiliary model's capacity. Although some methods designed for image understanding also combine lightweight query relevance with visual diversity for token selection Zhang et al. (2026); Li et al. (2026), they do not explicitly address how to allocate budgets across sparsely sampled video frames while preserving complementary query evidence within each frame.

These limitations highlight the need for a lightweight pre-LLM framework that estimates query relevance without auxiliary models or LLM attention, guiding both frame-level budget routing and intra-frame evidence selection to preserve informative visual context and complementary query-relevant evidence.

3 Method

3.1 Preliminaries

VideoLLM Architecture. Video large language models (VideoLLMs) Zhang et al. (2024); Li et al. (2024a); Bai et al. (2025) typically consist of three core components: **(1) Vision Encoder** extracts dense visual features $\mathcal{Z} = \{z_t\}_{t=1}^{F} \in \mathbb{R}^{F\times L\times D}$ from the raw video sequence



Figure 2: Visualizing different strategies for identifying query-relevant visual tokens. The color gradient indicates the degree of relevance, where cooler blue tones represent lower similarity and warmer red tones denote higher relevance.

$\mathcal{V} = \{v_t\}_{t=1}^{F} \in \mathbb{R}^{F\times H\times W\times 3}$, where F, L, and D denote the number of sampled frames, spatial patches per frame, and the visual feature dimension, respectively. **(2) Projector** serves as a cross-modal bridge mapping the visual embeddings into the LLM input embedding space. It transforms $\mathcal{Z}$ into LLM-readable visual tokens $\mathcal{X}_v = \{x_i^v\}_{i=1}^{F} \in \mathbb{R}^{F\times M\times D'}$, where $M \leq L$ accounts for spatial pooling, and D' matches the LLM's hidden dimension. **(3) LLM** processes the aligned visual tokens $\mathcal{X}_v$ alongside the text query tokens $\mathcal{X}_q = \{x_j^q\}_{j=1}^{N} \in \mathbb{R}^{N\times D'}$. Leveraging deep attention mechanisms, it performs cross-modal context fusion to autoregressively decode the natural language response $\mathcal{Y} = \{y_k\}_{k=1}$ via $y_k = \text{LLM}(\mathcal{X}_v, \mathcal{X}_q, \mathcal{Y}_{<k})$.

Token Compression for VideoLLMs. Token compression for VideoLLMs aims to accelerate inference by eliminating visual redundancy via token pruning or merging strategies. To this end, this process compresses a dense visual token sequence $\mathcal{X}_v$ into a drastically reduced sequence $\widetilde{\mathcal{X}}_v$ before or during LLM ingestion:

$$\widetilde{\mathcal{X}}_v = \Phi(\mathcal{X}_v), \quad \text{s.t.} \quad |\widetilde{\mathcal{X}}_v| \ll |\mathcal{X}_v|, \tag{1}$$

where $\Phi(\cdot)$ denotes the compression operator, and $|\cdot|$ measures the total token length. However, Eq. 1 formulates token compression as query-agnostic visual summarization, whereas video question answering requires prioritizing query-relevant evidence over generic visual representativeness.

3.2 Pre-LLM Query Relevance in the Aligned Embedding Space

To enable query-conditioned token compression, a reliable query-relevant signal is required. Building on prior studies of query-aware token selection in image understanding Zhang et al. (2026); Li et al. (2026), we investigate whether cosine similarity between projected visual tokens and embedded query tokens in the shared embedding space can capture meaningful query relevance both across and within sparsely sampled video frames. As in Figure 2, irrelevant frames tend to obtain low relevance scores, whereas visual regions corresponding to query-related concepts exhibit substantially higher similarity. More detailed analysis of identifying query-relevant visual tokens is provided in Appendix C. We therefore reformulate Eq. 1 as a query-conditioned compression formulation:

$$\widetilde{\mathcal{X}}_v = \Phi(\mathcal{X}_v, S(\mathcal{X}_v, \mathcal{X}_q)), \quad \text{s.t.} \quad |\widetilde{\mathcal{X}}_v| \ll |\mathcal{X}_v|, \tag{2}$$

where $S(\mathcal{X}_v, \mathcal{X}_q) = \{s(x_i^v, x_j^q)\}_{i,j}$ and $s(a,b) = \frac{<a,b>}{\|a\|_2\|b\|_2}$ denotes the cosine similarity between a and b, with $<\cdot,\cdot>$ representing the standard inner product, and $\|\cdot\|_2$ indicating the ℓ_2-norm.

3.3 EviQ: Query-Conditioned Evidence Selection

Although query relevance helps align visual token selection with task requirements, relying on relevance alone remains insufficient. Highly relevant tokens may repeatedly correspond to the same dominant query concepts, leading to redundant evidence while discarding complementary semantics and contextual cues needed for downstream reasoning. Thus, rather than simply retaining the tokens with the highest relevance scores, our goal is to construct a compact query-conditioned evidence set that provides comprehensive query coverage while preserving sufficient visual context.

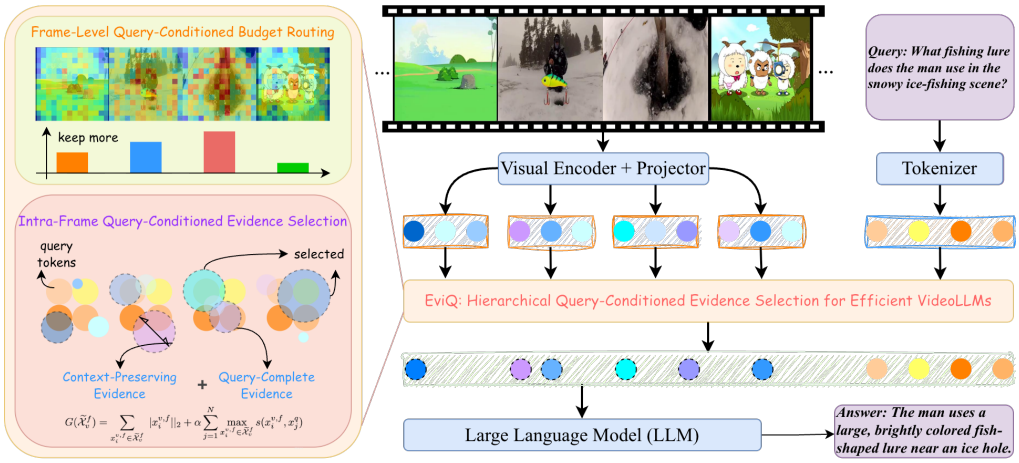


Figure 3: Overview of **EviQ**. Before LLM inference, **EviQ** first allocates the global visual-token budget across frames using aggregated visual–query similarities. Within each frame, it then greedily selects tokens according to their joint marginal contribution to visual informativeness and query-token coverage.

To this end, we propose **EviQ**, a coarse-to-fine evidence selection framework with two complementary stages. At the frame level, **EviQ** adaptively allocates the global token budget to frames that are more likely to contain query-relevant evidence. Given the allocated budget for each frame, it then performs intra-frame token selection by jointly preserving visually informative context and covering diverse query semantics. Through this two-stage process, **EviQ** produces a compact visual evidence set that allocates capacity toward query-relevant frames and promotes query-semantic coverage within each frame while preserving informative visual context under a fixed token budget. The overall framework is illustrated in Figure 3.

3.4 Frame-Level Query-Conditioned Budget Routing

Given a target retention ratio R, we set the global visual token budget to $K_{\text{global}} = FMR$, where F and M denote the number of sampled frames and tokens per frame, respectively. Since query-critical evidence is unevenly distributed across frames, **EviQ** allocates more tokens to query-relevant frames. We estimate frame relevance by averaging all pairwise visual–query token similarities:

$$s_f = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} s(x_i^{v,f}, x_j^q), \tag{3}$$

where $x_i^{v,f} \in \mathbb{R}^{D'}$ denotes the i-th visual token of the f-th frame. Rather than discarding low-relevance frames through hard thresholding, we standardize frame relevance scores and apply softmax with concentration coefficient τ to obtain a soft budget distribution:

$$\hat{s}_f = \frac{s_f - \mu_s}{\max(\sigma_s, 10^{-8})}, \qquad p_f = \frac{\exp(\hat{s}_f \cdot \tau)}{\sum_{j=1}^{F} \exp(\hat{s}_j \cdot \tau)}, \tag{4}$$

where μ_s and σ_s are the mean and population standard deviation of the frame relevance scores, respectively, and τ controls the sharpness of the resulting budget distribution. A larger τ concentrates more tokens on highly relevant frames, whereas $\tau = 0$ reduces to uniform allocation. The initial token budget allocated to the f-th frame is given by $k_f = \min(\lfloor p_f \cdot K_{\text{global}} \rfloor, M)$, and the remaining tokens are redistributed using the largest-remainder method to satisfy the global budget constraint.

3.5 Intra-Frame Query-Conditioned Evidence Selection

Given the allocated frame budget k_f, **EviQ** further selects the k_f visual tokens that best preserve the evidence required for answering the input query. Rather than independently ranking tokens according to query relevance, our objective is to construct a compact intra-frame evidence set that

preserves contextual cues while covering diverse query semantics within the frame. To this end, the intra-frame objective combines two complementary criteria: **Context-Preserving Evidence (CPE)**, which preserves informative scene context, and **Query-Complete Evidence (QCE)**, which promotes comprehensive coverage of the query semantics.

Context-Preserving Evidence (CPE). Query relevance alone is insufficient, since contextual regions may provide essential supporting evidence despite low query similarity. We therefore use the ℓ_2 norm of visual tokens as a lightweight proxy for visual informativeness Oyama et al. (2023); Sun et al. (2024), helping retain salient scene structures and supporting contextual cues.

Query-Complete Evidence (QCE). Given the visual evidence available in the current frame, the selected subset should cover as many query semantics as possible. We therefore evaluate tokens as a set by allowing only the best-matching visual token to contribute to each query token, naturally encouraging complementary coverage through diminishing returns.

By jointly considering contextual informativeness and frame-local query coverage, the quality of a selected subset $\widetilde{\mathcal{X}}_v^f$ within the f-th frame is defined as

$$G(\widetilde{\mathcal{X}}_v^f) = \sum_{x_i^{v,f} \in \widetilde{\mathcal{X}}_v^f} \|x_i^{v,f}\|_2 + \alpha \sum_{j=1}^{N} \max_{x_i^{v,f} \in \widetilde{\mathcal{X}}_v^f} s'(x_i^{v,f}, x_j^q), \tag{5}$$

where α controls the balance between the visual informativeness prior and the query-coverage gain, and $s' = \frac{s - s_{\min}}{s_{\max} - s_{\min} + 10^{-8}} \in [0, 1]$ is the min-max normalization of the cosine similarity within each frame. The intra-frame token selection problem can then be formulated as a cardinality-constrained maximization problem:

$$\max_{\widetilde{\mathcal{X}}_v^f \subseteq \mathcal{X}_v^f} G(\widetilde{\mathcal{X}}_v^f) \quad \text{s.t.} \quad |\widetilde{\mathcal{X}}_v^f| = k_f, \tag{6}$$

where $|\cdot|$ denotes set cardinality. Although finding the exact optimum is NP-hard, the objective G enjoys two key structural properties that enable efficient approximation.

Proposition 1. *The objective G in Eq. 5 is monotone and submodular.*

Proposition 1, proved in Appendix A.1, places our problem within the classical family of cardinality-constrained monotone submodular maximization problems Krause & Golovin (2014). We therefore adopt a greedy hill-climbing algorithm Nemhauser et al. (1978); Sha et al. (2016), whose approximation quality is characterized by the following corollary.

Corollary 2. *Under Proposition 1, the greedy algorithm produces a set $\widetilde{\mathcal{X}}_v^f$ that satisfies*

$$G(\widetilde{\mathcal{X}}_v^f) \;\geq\; \left(1 - \tfrac{1}{e}\right) \max_{\mathcal{X'}_v^f : |\mathcal{X'}_v^f| \leq k_f} G(\mathcal{X'}_v^f).$$

The full proof of Corollary 2 is provided in Appendix A.2. Thus, for each fixed frame budget k_f, greedy selection achieves the standard $(1 - 1/e)$ approximation guarantee with respect to the objective G.

4 Experiments

4.1 Experimental Settings

Benchmarks. We conduct evaluations on several multimodal video understanding benchmarks, including MVBench Li et al. (2024b), LongVideoBench Wu et al. (2024), MLVU Zhou et al. (2025), and VideoMME Fu et al. (2025a), using the LMMs-Eval Zhang et al. (2025a) evaluation framework. Further details are provided in Appendix D.

Implementation details. We evaluate the effectiveness of our method on multiple representative VideoLLMs, including LLaVA-OneVision Li et al. (2024a), LLaVA-Video Zhang et al. (2024), and Qwen3-VL Bai et al. (2025), to demonstrate its general applicability across different model architectures. Unless otherwise specified, we follow the official inference configuration of each backbone model and set $\tau = 0.5$ and $\alpha = 1.0$ for our method. Specifically, LLaVA-OneVision and Qwen3-VL uniformly sample 32 frames as input, whereas LLaVA-Video uniformly samples 64 frames. All experiments are conducted with the LMMs-Eval evaluation framework Zhang et al. (2025a) on a single NVIDIA H200 141GB GPU.

Table 1: Performance comparison of LLaVA-OV-7B with different methods across multiple benchmarks. All methods use the default 32-frame input setting with a retention ratio of 25%.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
LLaVA-OV-7B	58.3	56.4	63.0	69.9	56.6	48.8	58.4	59.0	100.0
DyCoke[CVPR25]	50.7	48.1	55.8	61.1	48.5	43.2	51.0	51.4	87.1
PruneVid[ACL25]	56.8	55.4	61.0	67.3	54.8	47.2	56.4	57.4	97.3
HoliTom[NIPS25]	**58.2**	56.3	63.0	69.0	56.0	46.4	57.1	58.7	99.4
CDPruner[NIPS25]	54.4	55.2	60.9	**69.7**	53.3	48.3	57.1	56.9	96.4
QuoTA[AAAI26]	57.2	56.7	61.6	68.2	56.4	46.6	57.1	58.2	98.6
FrameFusion[ICCV25]	56.0	54.6	61.6	68.2	55.6	48.6	57.5	57.4	97.3
FastVID[NIPS25]	56.5	56.1	60.8	69.4	**58.1**	47.2	58.3	57.9	98.2
VidCom^2[EMNLP25]	57.0	55.2	62.7	69.3	56.2	49.4	**58.4**	58.3	98.8
MMG-Vid[AAAI26]	55.0	53.5	61.3	67.1	54.7	47.5	56.4	56.6	95.8
FlashVID[ICLR26]	58.0	56.6	**63.2**	**69.7**	55.1	48.3	57.7	58.9	99.8
ForestPrune[CVPR26]	55.2	55.6	61.7	67.9	55.1	47.0	56.7	57.3	97.1
EviQ(ours)	57.7	**57.4**	63.0	68.7	55.2	**49.7**	57.8	**59.0**	**99.9**

Table 2: Performance comparison of LLaVA-Video-7B with different methods across multiple benchmarks. All methods use the default 64-frame input setting with a retention ratio of 25%.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
LLaVA-Video-7B	60.3	58.8	67.3	77.3	62.3	53.4	64.3	62.7	100.0
DyCoke[CVPR25]	50.8	53.0	56.9	65.8	53.6	48.9	56.1	54.2	86.3
HoliTom[NIPS25]	**58.4**	57.1	60.5	74.6	**62.3**	52.1	**63.0**	59.8	95.3
CDPruner[NIPS25]	56.2	56.1	60.7	72.9	60.1	50.7	61.2	58.6	93.4
QuoTA[AAAI26]	57.1	55.5	63.0	73.1	60.4	51.4	61.7	59.3	94.6
FrameFusion[ICCV25]	56.5	**57.7**	**65.9**	72.8	59.3	51.9	61.3	60.4	96.3
FastVID[NIPS25]	57.9	56.3	60.4	74.6	61.2	**52.9**	62.9	59.4	94.7
VidCom^2[EMNLP25]	57.0	56.7	58.7	73.0	61.7	50.0	61.6	58.5	93.3
MMG-Vid[AAAI26]	54.3	55.7	61.7	70.8	59.2	51.2	60.4	58.0	92.6
FlashVID[ICLR26]	57.9	56.7	62.2	**75.0**	61.2	51.8	62.7	59.9	95.4
EviQ(ours)	57.9	57.5	65.1	74.7	61.1	51.0	62.3	**60.7**	**96.8**

Compared baselines. We evaluate our method against a broad range of training-free token compression methods for VideoLLMs. These include similarity/uniqueness/diversity-based pre-LLM pruning methods, such as FastVID Shen et al. (2026), VidCom^2 Liu et al. (2025), MMG-Vid Ma et al. (2026), FlashVID Fan et al. (2026) and ForestPrune Ju et al. (2026); attention-based intra-LLM pruning methods, such as FrameFusion Fu et al. (2025b); and multi-stage similarity&attention-based pruning methods, such as DyCoke Tao et al. (2025), PruneVid Huang et al. (2025), and HoliTom Shao et al. (2026). We also compare our method with CDPruner Zhang et al. (2026) and QuoTA Luo et al. (2026), two query-driven pruning methods for visual large language models (VLLMs). The **best result** for each column is highlighted in bold, and the second-best result is underlined in each table. More details are provided in Appendix E.

4.2 Main Results

Effectiveness on Accuracy. Tables 1, 2, and 3 report the performance comparisons on LLaVA-OV-7B, LLaVA-Video-7B, and Qwen3-VL-8B-Instruct, respectively, across diverse video understanding benchmarks. **(i) State-of-the-Art Overall Performance.** As shown in these tables, **EviQ** consistently achieves the best overall accuracy and state-of-the-art performance among training-free token compression methods, obtaining the highest average scores across all evaluated models under a 25% retention ratio. Notably, on LLaVA-OV-7B, **EviQ** nearly preserves the full-model performance with only 25% visual tokens, achieving a relative accuracy of **99.9%**. **(ii) Strong Long-Video Understanding Capability. EviQ** also demonstrates clear advantages on long-video understanding tasks, achieving best or second-best performance across LongVideoBench and MLVU. These results indicate that a query-aware strategy is more effective than relying solely on visual cues for retaining query-relevant visual information, especially when the critical evidence is temporally distributed and localized. **(iii) Robustness under Limited Token Budgets.** Furthermore, our method remains robust even under more aggressive compression. When the retention ratio is reduced to 15% on Qwen3-VL, **EviQ** still achieves the highest average score and retains 96.0% of the original model performance. This demonstrates that **EviQ** can effectively identify compact yet informative

Table 3: Performance comparison of Qwen3-VL-8B-Instruct with different methods across multiple benchmarks. All methods use the default 32-frame input setting.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
Qwen3-VL-8B-Instruct	68.6	60.2	63.4	76.0	60.0	56.9	64.3	64.1	100.0
Retention Ratio=25%									
HoliTom[NIPS25]	63.0	56.8	61.2	71.4	54.6	53.1	59.7	60.2	93.8
CDPruner[NIPS25]	67.0	58.3	61.6	73.7	58.6	53.8	62.0	62.2	97.0
FastVID[NIPS25]	67.3	58.6	60.7	72.0	56.8	52.6	60.5	61.8	96.4
VidCom^2[EMNLP25]	67.0	57.9	60.6	72.0	**59.3**	**56.2**	**62.5**	62.0	96.7
MMG-Vid[AAAI26]	66.9	58.3	60.6	70.2	55.8	53.0	59.7	61.4	95.7
FlashVID[ICLR26]	67.5	58.7	61.6	**74.4**	58.3	53.8	62.1	62.5	97.4
EviQ(ours)	**67.9**	**59.8**	**61.7**	71.9	57.6	53.1	60.9	**62.6**	**97.7**
Retention Ratio=15%									
HoliTom[NIPS25]	60.0	55.7	59.5	68.9	54.3	51.6	58.3	58.4	91.1
CDPruner[NIPS25]	63.8	56.3	60.6	70.9	**58.3**	**54.3**	**61.2**	60.5	94.3
FastVID[NIPS25]	66.1	57.0	58.1	69.6	53.9	51.0	58.1	59.8	93.3
VidCom^2[EMNLP25]	64.2	55.9	57.6	68.9	56.3	53.2	59.4	59.3	92.4
MMG-Vid[AAAI26]	65.2	58.0	59.3	66.9	54.0	51.2	57.4	60.0	93.5
FlashVID[ICLR26]	**66.5**	57.7	59.9	**72.4**	57.2	54.0	**61.2**	61.3	95.6
EviQ(ours)	66.2	**59.2**	**60.8**	69.1	56.3	54.1	59.9	**61.5**	**96.0**

Table 4: Efficiency comparison of LLaVA-OV-7B with different methods across a total of four benchmarks. All methods use the default 32-frame input setting with a retention ratio of 25%.

Method	Encode Latency(ms)	Prefill Latency(ms)	Overhead Latency(ms)	Total Latency(ms)	GPU Average Memory(MB)	Performance
LLaVA-OV-7B	175.2	255.7	0.0	462.4	22086	59.0
HoliTom[NIPS25]	225.5	68.1	49.9	325.1(×1.42)	20178(↓8.6%)	58.7(↓**0.3**)
CDPruner[NIPS25]	383.7	68.1	207.9	483.0(×0.96)	20168(↓8.7%)	56.9(↓2.1)
QuoTA[AAAI26]	2874.1	67.6	2696.8	2970.7(×0.16)	27028(↑22.4%)	58.2(↓0.8)
FastVID[NIPS25]	180.5	67.8	5.0	280.0(×**1.65**)	20178(↓8.6%)	57.9(↓1.1)
VidCom^2[EMNLP25]	177.7	67.4	2.2	276.4(×**1.67**)	20168(↓8.7%)	58.3(↓0.7)
MMG-Vid[AAAI26]	196.3	67.7	20.7	295.9(×1.56)	20168(↓8.7%)	56.6(↓2.4)
FlashVID[ICLR26]	216.1	67.9	40.7	316.1(×1.46)	20182(↓8.6%)	58.9(↓**0.1**)
EviQ(ours)	191.3	67.6	15.7	291.1(×**1.59**)	20168(↓8.7%)	59.0 (★)

token subsets under constrained budgets, while consistently maintaining strong average performance across models and evaluation suites.

Effectiveness on Efficiency. Table 4 compares the efficiency of LLaVA-OV-7B equipped with different training-free token pruning methods, where total latency is the sum of encode, prefill, and decode latency, and decode latency remains about 32 ms for all methods, while the reported overhead measures the time spent on token pruning included in the encode or prefill latency. Existing methods typically face an accuracy-efficiency trade-off, where lightweight approaches such as FastVID and VidCom^2 trade noticeable performance drops for higher throughput, while accuracy-preserving methods such as HoliTom and FlashVID incur large overhead that limits overall acceleration. By comparison, **EviQ** achieves a more favorable balance between inference efficiency and model performance, achieving a 1.59× speedup with only 5.4% additional overhead. Meanwhile, **EviQ** preserves nearly all of the original accuracy, thereby demonstrating a superior accuracy-efficiency trade-off compared to existing pruning methods. Moreover, our method is orthogonal to FlashAttention Dao et al. (2022); Dao (2024) and consistently reduces average GPU memory usage by 8.7%, further confirming its practical efficiency.

Generalization Ability. We further evaluate the generalization ability of our method across model architectures and video sampling strategies. First, as shown in Tables 1, 2 and 3, **EviQ** consistently achieves the best average performance, demonstrating that it is not tailored to a specific model architecture or inference configuration but generalizes well across different VideoLLMs and retention ratios. Second, as shown in Table 5, we evaluate our method on LongVideoBench with LLaVA-OneVision under three video sampling strategies, including uniform sampling, AKS Tang et al. (2025), and MDP3 Sun et al. (2025), and report accuracy (%) for video-duration ranges of *(0,15]*, *(15,60]*, *(60,600]*, and *(600,3600]* seconds. **EviQ** consistently achieves the best overall performance among visual token pruning methods and maintains strong performance on the *3600s* split across different sampling settings, indicating its ability to preserve task-relevant visual tokens for long-duration videos. These results further highlight its ability to robustly generalize across diverse VideoLLMs and its compatibility with different sampling strategies.

Table 5: Generalization performance under different sampling methods on LongVideoBench with LLaVA-OneVision-7B at a retention ratio of 25%.

Sample Method	Prune Method	LongVideoBench				
		15s	*60s*	*600s*	*3600s*	**Overall**
Uniform	-	70.4	66.9	57.0	48.6	56.4
	HoliTom[NIPS25]	67.7	69.8	57.3	47.7	56.3
	FlashVID[ICLR26]	69.3	66.4	56.5	49.5	56.6
	EviQ(ours)	71.4	70.4	55.6	50.4	**57.4**
AKS [CVPR25]	-	67.9	69.9	57.9	57.6	60.6
	HoliTom[NIPS25]	70.5	67.3	56.9	48.6	56.5
	FlashVID[ICLR26]	72.0	65.4	56.7	50.2	57.1
	EviQ(ours)	72.0	67.3	57.4	50.5	**57.7**
MDP3 [ICCV25]	-	71.5	69.3	59.3	51.9	59.1
	HoliTom[NIPS25]	69.9	69.9	58.6	50.9	58.3
	FlashVID[ICLR26]	69.4	68.6	58.8	50.5	58.0
	EviQ(ours)	71.5	67.3	59.5	51.9	**58.9**

Table 6: Ablation study on different modules in **EviQ**. QCBR, CPE, and QCE denote Query-Conditioned Budget Routing, Context-Preserving Evidence, and Query-Complete Evidence, respectively.

Method	MVBench	MLVU	VideoMME
LLaVA-OV-7B, Retention Ratio=25%			
Vanilla	58.3	63.0	58.4
CPE	57.5	62.6	56.1
QCE	51.0	55.2	52.7
CPE + QCE	57.5	62.7	56.6
QCBR + CPE	57.7	62.4	56.7
QCBR + QCE	54.5	61.2	57.0
EviQ(ours)	57.7	63.0	57.8

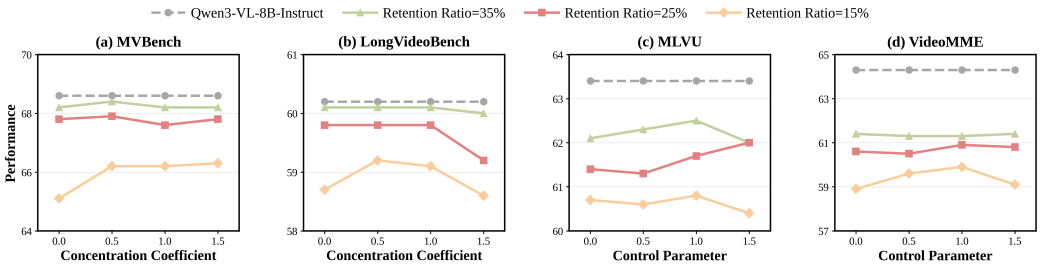


Figure 4: Hyperparameter sensitivity analysis for concentration coefficient τ in Eq.4 and control parameter α in Eq.5 under different retention ratios on Qwen3-VL-8B across four benchmarks.

4.3 Ablation Study and Analysis

Ablation on EviQ components. Table 6 presents the ablation study of different **EviQ** components on LLaVA-OV-7B under a retention ratio of 25%: **(i) Context-Preserving Evidence (CPE)** preserves visually informative tokens and provides a stable pruning baseline across different benchmarks. **(ii) Query-Complete Evidence (QCE)** encourages the selected tokens to better cover query-relevant visual evidence, complementing CPE by improving task relevance. **(iii) Query-Conditioned Budget Routing (QCBR)** adaptively allocates larger token budgets to more query-relevant frames and enables frame-wise greedy selection, reducing selection time by 61.6% versus global greedy selection over all frames' tokens. The complete **EviQ** achieves the best overall results, demonstrating that the three components jointly provide the most robust performance.

Ablation on hyperparameters. To evaluate the sensitivity of **EviQ**, we vary the concentration coefficient τ on MVBench and LongVideoBench, and the balance parameter α on MLVU and VideoMME, as shown in Figure 4. **(i) The concentration coefficient** τ controls the sharpness of frame-level budget routing. Smaller values produce a more uniform distribution, while larger values allocate more tokens to frames with higher query relevance. When $\tau = 0$, routing degenerates to uniform allocation, disabling frame-level query-conditioned budget routing. Across retention ratios, varying τ causes only marginal fluctuations on MVBench and LongVideoBench. **(ii) The control parameter** α balances CPE and QCE in intra-frame selection. Larger values emphasize query-semantic coverage, while smaller values rely more on visually informative context. When $\alpha = 0$, selection reduces to CPE-only token selection. Varying α also leads to limited fluctuations on MLVU and VideoMME. We therefore set $\tau = 0.5$ and $\alpha = 1.0$ by default.

5 Conclusion

In this paper, we present **EviQ**, a training-free pre-LLM token pruning framework for efficient VideoLLM inference. Using cosine similarity between aligned visual and query tokens as a lightweight relevance signal, it builds a compact query-conditioned evidence set through frame-level budget routing and intra-frame token selection. By preserving informative visual context while promoting complementary query coverage, **EviQ** retains reasoning-relevant evidence under a limited token budget. Across multiple VideoLLMs, benchmarks, and retention ratios, it consistently achieves SOTA accuracy-efficiency trade-offs among training-free token compression methods. It thus highlights pre-LLM evidence construction as a promising path toward efficient VideoLLMs, motivating future work on more accurately identifying key evidence and reasoning over it more effectively.

AI USE STATEMENT

In this work, we used generative AI tools for several tasks with required disclosure. Specifically, we used them to assist in the writing of proofs, to assist with translation, to support qualitative and thematic data analysis, and to interpret results. We did not use generative AI tools for the other tasks with required disclosure, namely developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, implementing methods, and designing or providing feedback on research methodology or experiments. The remaining tasks with required disclosure, namely generating synthetic datasets and cleaning or reformatting datasets, are not applicable to this work.

Additionally, we used generative AI tools for tasks with recommended disclosure, namely creating or modifying scientific figures or images, drafting parts of the research paper, summarizing or analyzing existing literature, sourcing and searching for information, editing the research paper to improve readability, identifying relevant literature, and proposing a title or keywords for the research paper. We have reviewed all AI-assisted work. All AI-assisted proofs, analyses, and result interpretations were independently verified by the authors, who also reviewed and, as needed, revised all other AI-assisted text, figures, and cited literature to ensure their accuracy. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jingfeng Chen, Jiawen Qian, Wendi Deng, Yinuo Guo, Jiaqi Yu, Sicong Leng, Raghuveer Thirukovalluru, and Bhuwan Dhingra. Lddr: Linear-dpp-based dynamic-resolution frame sampling for video mllms. *arXiv preprint arXiv:2605.11477*, 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. Flashvid: Efficient video large language models via training-free tree-based spatiotemporal token merging. *arXiv preprint arXiv:2602.08024*, 2026.

Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4): 634–652, 1998.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025a.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22654–22663, 2025b.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Shaobo Ju, Baiyang Song, Tao Chen, Jiapeng Zhang, Qiong Wu, Chao Chang, Huaixi Wang, Yiyi Zhou, and Rongrong Ji. Forestprune: High-ratio visual token compression for video multimodal large language models via spatial-temporal forest modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8326–8336, June 2026.

Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability*, 3(71-104): 3, 2014.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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

Xu Li, Yi Zheng, Mengyang Zhao, Yuxuan Liang, Zhe Liu, Rui Zhu, Xiaolei Chen, Wei Zhou, Baoquan Zhao, and Juncen Guo. Crisp: Pre-llm yet text-driven visual token pruning for efficient lvlm inference. *arXiv preprint arXiv:2607.16326*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1910–1924, 2025.

Yongdong Luo, Wang Chen, Weizhong Huang, Shukang Yin, Haojia Lin, Jinfa Huang, Chaoyou Fu, Jiayi Ji, Xiawu Zheng, and Jiebo Luo. Quota: Query-oriented token assignment via cot query decouple for long video comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24160–24168, 2026.

Junpeng Ma, Qizhe Zhang, Ming Lu, Zhibin Wang, Qiang Zhou, Jun Song, and Shanghang Zhang. Mmg-vid: Maximizing marginal gains at segment-level and token-level for efficient video llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24253–24261, 2026.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. Norm of word embedding encodes information gain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2108–2130, 2023.

David Picard. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Chaofeng Sha, Xiaowei Wu, and Junyu Niu. A framework for recommending relevant and diverse items. In *IJCAI*, volume 16, pp. 3868–3874, 2016.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *Advances in Neural Information Processing Systems*, 38:135547–135570, 2026.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, et al. Fastvid: Dynamic density pruning for fast video large language models. *Advances in Neural Information Processing Systems*, 38:123553–123581, 2026.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24090–24101, 2025.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 4942–4964, 2024.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29118–29128, 2025.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, 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, 2025a.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *Advances in Neural Information Processing Systems*, 38:25438–25468, 2026.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

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

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. *arXiv preprint arXiv:2510.27280*, 2025.

A Theoretical Analysis of the Token Selection Objective

To complement the algorithmic description in the main text, we provide a self-contained analysis of the intra-frame token selection objective in Eq. 5. For notational simplicity, we omit the frame index f and the modality index v, and rewrite the objective as

$$G(\widetilde{\mathcal{X}}) = \underbrace{\sum_{x\in\widetilde{\mathcal{X}}} \|x\|_2}_{G_1(\widetilde{\mathcal{X}})} + \alpha \underbrace{\sum_{j=1}^{N} \max_{x\in\widetilde{\mathcal{X}}} s'(x, x_j^q)}_{G_2(\widetilde{\mathcal{X}})}, \tag{7}$$

where $\alpha \geq 0$ and $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$. By construction, the rescaled similarity satisfies $s'(x, x_j^q) = (s(x, x_j^q) - s_{\text{min}})/(s_{\text{max}} - s_{\text{min}} + 10^{-8}) \in [0, 1]$ for every visual token x and query token x_j^q. We adopt the convention $\max_{x\in\emptyset} s'(x, x_j^q) = 0$, which gives $G(\emptyset) = 0$.

We first recall the structural properties used in the analysis.

Definition 3 (Monotonicity)**.** A set function $f : 2^V \to \mathbb{R}$ is *monotone non-decreasing* if $f(A) \leq f(B)$ for every $A \subseteq B \subseteq V$.

Definition 4 (Submodularity)**.** A set function $f : 2^V \to \mathbb{R}$ is *submodular* if, for every $A \subseteq B \subseteq V$ and every $e \in V \setminus B$,

$$f(A \cup \{e\}) - f(A) \;\geq\; f(B \cup \{e\}) - f(B). \tag{8}$$

This is the diminishing-returns property: the marginal gain of adding an element cannot increase as the conditioning set grows.

Definition 5 (Modularity)**.** A set function $f : 2^V \to \mathbb{R}$ is *modular* if the marginal gain of each element is independent of the conditioning set, i.e.,

$$f(A \cup \{e\}) - f(A) \;=\; f(B \cup \{e\}) - f(B), \tag{9}$$

for every $A \subseteq B \subseteq V$ and every $e \in V \setminus B$. Equivalently, $f(S) = f(\emptyset) + \sum_{e\in S} w_e$ for some weights $\{w_e\}_{e\in V}$. Thus, every modular function is submodular.

A.1 Monotonicity and Submodularity of G

Proof of Proposition 1. We analyze the two components of G separately.

Step 1: G_1 is modular, monotone, and submodular. For any $A \subseteq \mathcal{X}$ and any $e \in \mathcal{X} \setminus A$,

$$G_1(A \cup \{e\}) - G_1(A) \;=\; \|e\|_2 \;\geq\; 0. \tag{10}$$

The marginal gain is non-negative and independent of A. Therefore, G_1 is monotone non-decreasing and modular. Since modularity is a special case of submodularity, G_1 is also submodular.

Step 2: G_2 is monotone. For each query token x_j^q, define $g_j(S) \triangleq \max_{x\in S} s'(x, x_j^q)$, with $g_j(\emptyset) = 0$. For any $A \subseteq B \subseteq \mathcal{X}$, the maximum in $g_j(B)$ is taken over a superset of the one in $g_j(A)$. Since $s'(\cdot, \cdot) \geq 0$, this also holds when $A = \emptyset$. Hence $g_j(A) \leq g_j(B)$, and summing over $j = 1, \ldots, N$ gives $G_2(A) \leq G_2(B)$.

Step 3: G_2 is submodular. Because submodularity is closed under non-negative linear combinations, it is sufficient to show that each g_j is submodular. Fix a query token x_j^q, let $A \subseteq B \subseteq \mathcal{X}$, and take any $e \in \mathcal{X} \setminus B$. Write $s_e \triangleq s'(e, x_j^q)$. There are three cases:

- If $s_e \leq g_j(A)$, then $s_e \leq g_j(B)$ as well, and adding e changes neither maximum. Both marginal gains are zero.
- If $g_j(A) < s_e \leq g_j(B)$, then e improves the maximum for A but not for B. Thus, $g_j(A \cup \{e\}) - g_j(A) = s_e - g_j(A) > 0$, while $g_j(B \cup \{e\}) - g_j(B) = 0$.
- If $s_e > g_j(B)$, then e becomes the new maximizer for both sets. The two marginal gains are $s_e - g_j(A)$ and $s_e - g_j(B)$, respectively. Since $g_j(A) \leq g_j(B)$, the former is no smaller than the latter.

In all cases,

$$g_j(A \cup \{e\}) - g_j(A) \;\geq\; g_j(B \cup \{e\}) - g_j(B), \tag{11}$$

so g_j is submodular. Summing over all query tokens shows that G_2 is submodular.

Step 4: Combining the components. Since $\alpha \geq 0$, and both G_1 and G_2 are monotone and submodular, their non-negative linear combination $G = G_1 + \alpha G_2$ is also monotone and submodular. This completes the proof of Proposition 1. □

A.2 $(1 - 1/e)$ Approximation Guarantee

We now prove the $(1 - 1/e)$ approximation guarantee for the greedy hill-climbing procedure used in **EviQ**. The procedure starts from $S_0 = \emptyset$ and, at step $i + 1$, adds the visual token with the largest marginal gain:

$$e_{i+1} \;\in\; \arg\max_{e \in \mathcal{X} \setminus S_i} \big[\, G(S_i \cup \{e\}) - G(S_i) \,\big], \qquad S_{i+1} = S_i \cup \{e_{i+1}\}. \tag{12}$$

After k_f steps, the algorithm returns a set S_{k_f} with $|S_{k_f}| = k_f$. Let $S^\star \in \arg\max_{S \subseteq \mathcal{X},\, |S| = k_f} G(S)$ denote an optimal solution, and write $G^\star \triangleq G(S^\star)$. We assume $k_f \geq 1$.

Proof of Corollary 2. The proof follows the classical analysis of Nemhauser, Wolsey, and Fisher Nemhauser et al. (1978), specialized to our objective G, which is monotone and submodular by Proposition 1.

Enumerate $S^\star = \{o_1, \ldots, o_{k_f}\}$ in an arbitrary order. Since G is monotone and $S^\star \subseteq S_i \cup S^\star$, we have

$$G^\star \;\leq\; G(S_i \cup S^\star). \tag{13}$$

Expanding the right-hand side relative to S_i gives

$$G(S_i \cup S^\star) - G(S_i) = \sum_{l=1}^{k_f} \Big[G\big(S_i \cup \{o_1, \ldots, o_l\}\big) - G\big(S_i \cup \{o_1, \ldots, o_{l-1}\}\big) \Big]. \tag{14}$$

Each summand is the marginal gain of adding o_l after the set $S_i \cup \{o_1, \ldots, o_{l-1}\}$ has already been included. By submodularity, this marginal gain is no larger than the marginal gain of adding o_l directly to the smaller set S_i:

$$G\big(S_i \cup \{o_1, \ldots, o_l\}\big) - G\big(S_i \cup \{o_1, \ldots, o_{l-1}\}\big) \;\leq\; G(S_i \cup \{o_l\}) - G(S_i). \tag{15}$$

If $o_l \in S_i$, both sides are zero; otherwise, this is exactly the diminishing-returns property. Combining this inequality with Eq. 13 yields

$$G^\star - G(S_i) \;\leq\; \sum_{l=1}^{k_f} \Big[\, G(S_i \cup \{o_l\}) - G(S_i) \,\Big]. \tag{16}$$

At step $i+1$, the greedy rule chooses the element with the largest marginal gain among all candidates in $\mathcal{X} \setminus S_i$. Therefore, for every $o_l \notin S_i$,

$$G(S_i \cup \{o_l\}) - G(S_i) \;\leq\; G(S_{i+1}) - G(S_i). \tag{17}$$

For any $o_l \in S_i$, the left-hand side is zero and the same inequality still holds because G is monotone. Substituting into Eq. 16, we obtain

$$G^\star - G(S_i) \;\leq\; k_f \big[\, G(S_{i+1}) - G(S_i) \,\big]. \tag{18}$$

Equivalently,

$$G(S_{i+1}) - G(S_i) \;\geq\; \frac{1}{k_f} \big[\, G^\star - G(S_i) \,\big]. \tag{19}$$

Rearranging gives the residual recursion

$$G^\star - G(S_{i+1}) \;\leq\; \left(1 - \frac{1}{k_f}\right) \big[\, G^\star - G(S_i) \,\big]. \tag{20}$$

Iterating Eq. 20 for $i = 0, \ldots, k_f - 1$ and using $G(S_0) = G(\emptyset) = 0$ gives

$$G^\star - G(S_{k_f}) \;\le\; \left(1 - \frac{1}{k_f}\right)^{k_f} G^\star \;\le\; \frac{1}{e} G^\star, \tag{21}$$

where the last inequality follows from $(1 - 1/k)^k \le 1/e$ for all integers $k \ge 1$. Hence

$$G(S_{k_f}) \;\ge\; \left(1 - \frac{1}{e}\right) G^\star. \tag{22}$$

This proves Corollary 2. □

The approximation ratio $1 - 1/e \approx 0.632$ is tight for the general class of monotone submodular maximization problems under a cardinality constraint: a strictly better polynomial-time guarantee would imply a better-than- $(1 - 1/e)$ approximation for maximum coverage, which is impossible unless P = NP Feige (1998). In practice, the greedy solution is often much closer to the optimum than this worst-case guarantee suggests, consistent with the strong empirical performance of **EviQ**.

B The Algorithm Framework for EviQ

Algorithm 1 EviQ

Input: visual tokens $\mathcal{X}_v = \{x_i^{v,f}\}_{i,f}$, query tokens $\mathcal{X}_q = \{x_j^q\}_{j=1}^N$, retention ratio R, concentration coefficient τ, balancing coefficient α

 # Stage 1: Frame-Level Query-Conditioned Budget Routing.

1. compute cosine similarity $s(x_i^{v,f}, x_j^q)$
2. compute frame-level score s_f and normalized version p_f ▷ Eq.3 & Eq.4
3. compute budget allocation k_f based on largest-remainder method

 # Stage 2: Intra-Frame Query-Conditioned Evidence Selection.

rescale similarity with min-max normalization: $s' = (s - s_{\min})/(s_{\max} - s_{\min} + 10^{-8}) \in [0, 1]$

for $f = 1$ **to** F **do**
  $\widetilde{\mathcal{X}}_v^f \leftarrow \emptyset$
  while $|\widetilde{\mathcal{X}}_v^f| < k_f$ **do**
    $e^\star \in \arg\max_{e \in \mathcal{X}_v^f \setminus \widetilde{\mathcal{X}}_v^f} \big[\, G(\widetilde{\mathcal{X}}_v^f \cup \{e\}) - G(\widetilde{\mathcal{X}}_v^f) \,\big]$ ▷ Eq.5
    $\widetilde{\mathcal{X}}_v^f \leftarrow \widetilde{\mathcal{X}}_v^f \cup \{e^\star\}$
  end while
end for
$\widetilde{\mathcal{X}}_v = \bigcup_{f=1}^F \widetilde{\mathcal{X}}_v^f$

Output: pruned visual tokens $\widetilde{\mathcal{X}}_v$

C Analysis of Visual Token Importance Signals

Video and multimodal LLMs encode each frame into hundreds of visual tokens, producing highly redundant sequences whose length dominates the cost of the language model. Token pruning alleviates this by retaining only a subset of visual tokens, but its effectiveness hinges on a single question of what signal should be used to decide which tokens matter. The pruning mechanism is secondary. What matters is the importance signal, which determines both how much computation can be saved and whether the tokens that actually carry the answer are preserved. Below we contrast the two signals commonly used in prior work, namely **CLS attention** and **shallow-layer LLM attention**, with the **query-relevance** signal we adopt, and show that the latter is advantageous along both the efficiency and the effectiveness axes.

CLS attention, derived from the `[CLS]` token of the vision encoder, measures generic visual saliency and is entirely query-agnostic. It scores a token by how visually prominent it is, not by whether it is relevant to what the user is asking. As a result, it tends to preserve tokens that are salient but irrelevant while discarding tokens that are visually plain yet essential to the answer. This failure mode is exactly what Figure 5 illustrates, where CLS-based scores are drawn toward the

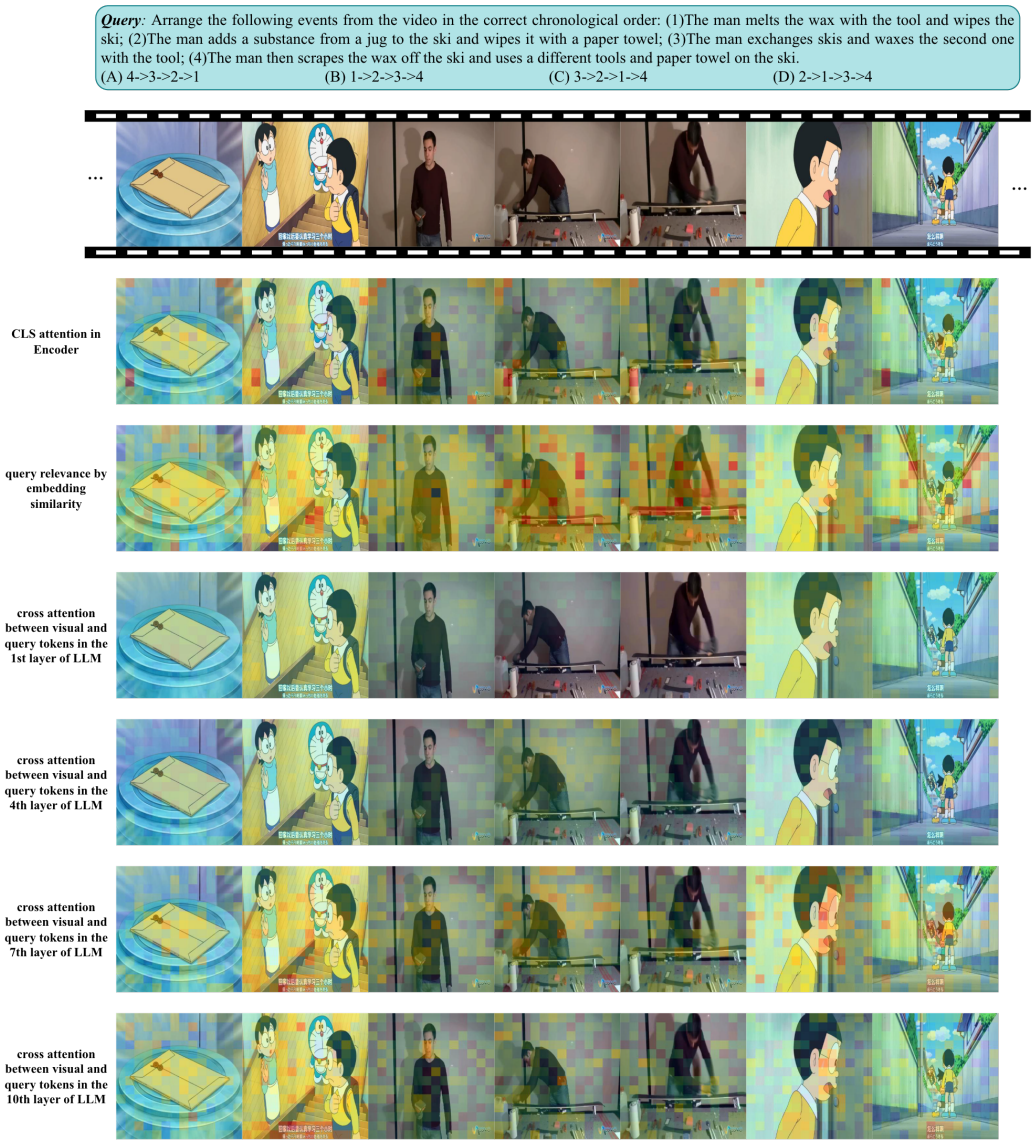


Figure 5: Comparison of visual token importance estimation methods. The color gradient indicates the degree of relevance, where cooler blue tones represent lower similarity and warmer red tones denote higher relevance.

vivid, high-contrast cartoon frames, even though only the plainer real-footage frames of the man waxing the ski are relevant to the temporal-ordering query.

Shallow-layer LLM attention scores tokens using the attention weights of the first few decoder layers. This introduces two problems at once. **(i)** Obtaining these scores requires running part of the LLM forward pass, so pruning happens inside the language model rather than at its entrance, and the layers consumed before pruning cannot be saved, which limits the achievable speedup. **(ii)** Early-layer attention is not yet well aligned with semantics and is corrupted by well-documented artifacts such as attention sinks and positional bias, making the signal noisy and unreliable. Shallow-layer attention therefore loses both on efficiency and on the quality of the signal it provides.

Query-relevance addresses both shortcomings simultaneously. Because relevance between visual tokens and the query is computed before the LLM forward pass, pruning occurs at the very entrance to the language model, so the discarded tokens never enter its most expensive layers, yielding the largest and cleanest speedup. At the same time, the signal directly measures alignment with the

actual query, so the retained tokens are precisely those needed to answer it, and salient-but-irrelevant content is naturally filtered out. These two advantages map cleanly onto the baselines. **(i)** Relative to shallow-layer attention, query-relevance wins on both axes, since it prunes pre-LLM rather than mid-LLM and its relevance signal is clean and directly query-aligned. **(ii)** Relative to CLS attention, which is also pre-LLM and thus comparable in efficiency, it wins on effectiveness, since being query-aware lets it avoid the salient-but-irrelevant trap that CLS cannot escape.

Figure 5 makes this contrast visually explicit. Query-relevance concentrates its importance mass on the real-footage frames that the query concerns, whereas CLS and shallow-layer attention are either diffuse or misdirected toward the irrelevant cartoon frames. This qualitative behavior is consistent with the quantitative gains reported in the main paper and supports our choice of a pre-LLM, query-relevance signal for token selection.

D Evaluation Benchmarks

VideoMME. VideoMME Fu et al. (2025a) is a large-scale benchmark for evaluating the video understanding ability of VLLMs. It includes 900 videos collected from six domains and 30 subcategories, with video lengths ranging from 11 seconds to 1 hour. The benchmark provides high-quality human annotations and contains 2,700 multiple-choice question-answer pairs.

LongVideoBench. LongVideoBench Wu et al. (2024) focuses on long-context video-language understanding. It consists of 3,763 videos, whose durations span from 8 seconds to 1 hour, and provides 6,678 human-annotated multiple-choice questions. The benchmark introduces a referring-reasoning setting, where models are required to locate the multimodal evidence referred to by the question and reason over it. Its questions cover 17 fine-grained categories at both perception and relation levels.

MVBench. MVBench Li et al. (2024b) is designed to evaluate temporal understanding in multi-modal video tasks. Unlike image-centric benchmarks, it emphasizes dynamic video reasoning and includes 20 carefully constructed tasks. These tasks are built through static-to-dynamic transformation and require models to capture temporal relations, actions, and event-level changes.

MLVU. MLVU Zhou et al. (2025) is a long-video understanding benchmark covering nine task types, including both multiple-choice and free-form generation tasks. Its videos range from 3 minutes to 2 hours, requiring models to perform event retrieval, temporal ordering, plot understanding, and video summarization. MLVU therefore evaluates both fine-grained visual perception and long-range temporal comprehension.

E Comparison Baselines

FastVID Shen et al. (2026) is a training-free dynamic density pruning framework for efficient VideoLLM inference. It first partitions a video into temporally coherent segments according to frame-level similarity and adaptively allocates token budgets across segments based on their information density. Within each segment, FastVID performs density-aware spatiotemporal pruning that discards locally redundant tokens while retaining tokens that capture salient motion or appearance changes, thereby preserving both the global temporal structure and the essential visual details of the video.

VidCom^2 Liu et al. (2025) is a plug-and-play video token compression method that adaptively adjusts the compression ratio according to frame-level uniqueness. It first estimates each frame's distinctiveness relative to its temporal neighbors and then assigns a lower compression ratio to more distinctive frames and a higher compression ratio to redundant ones, so that the overall token budget is concentrated on informative frames. This content-adaptive scheduling effectively reduces visual token redundancy while preserving the discriminative information needed for downstream video understanding.

MMG-Vid Ma et al. (2026) prunes visual tokens by maximizing marginal gains at both the segment and token levels. At the segment level, it partitions the video based on frame similarity and dynamically distributes the token budget across segments according to their marginal contribution to the overall representation. At the token level, it greedily selects representative tokens by jointly considering inter-frame uniqueness and intra-frame diversity, ensuring that each retained token brings non-redundant information to the already selected subset.

FlashVID Fan et al. (2026) is a training-free spatiotemporal token merging framework for efficient VideoLLMs. It first identifies a set of representative tokens using complementary attention- and diversity-based cues, which together highlight tokens that are both task-relevant and semantically heterogeneous. The remaining tokens are then progressively absorbed into these representatives through a tree-based spatiotemporal merging scheme, which removes fine-grained redundancy across neighboring frames while keeping the computational overhead negligible.

ForestPrune Ju et al. (2026) performs high-ratio visual token compression through spatiotemporal forest modeling. It organizes visual tokens across frames into a forest of trees under joint semantic, spatial, and temporal constraints, where each tree groups tokens that are mutually consistent along these three dimensions. Pruning decisions are then made hierarchically according to the importance of individual token trees and nodes, allowing aggressive token reduction while still retaining a structurally coherent subset of visual evidence.

FrameFusion Fu et al. (2025b) reduces video tokens by combining similarity-based merging with importance-based pruning in a single pipeline. It builds on the observation that high visual token similarities decrease substantially in deeper LLM layers, motivating the removal of similarity-based redundancy early in inference. Accordingly, FrameFusion merges highly similar tokens across successive early layers and then prunes low-importance tokens in deeper layers, achieving substantial token reduction with only limited degradation in downstream performance.

DyCoke Tao et al. (2025) is a dynamic token compression method designed for fast VideoLLM inference. It first applies a lightweight temporal token compression step that merges redundant tokens across adjacent frames before they enter the LLM. During decoding, DyCoke further performs dynamic KV cache reduction that adaptively retains different critical visual tokens at different decoding steps, so that the active context size remains small while the model can still attend to step-specific evidence in the video.

PruneVid Huang et al. (2025) is a training-free pruning framework that reduces video redundancy by merging temporally static and spatially similar tokens. It first collapses tokens that remain stable across consecutive frames and aggregates tokens that share similar spatial semantics, yielding a compact yet content-preserving token set. PruneVid then leverages the LLM's inherent query-aware reasoning ability to further filter this set, selectively discarding visual tokens that are weakly related to the input question and concentrating the remaining budget on query-relevant evidence.

HoliTom Shao et al. (2026) introduces a holistic token merging framework that unifies outer-LLM and inner-LLM compression (i.e., compression performed before and inside the language model). Outside the LLM, it conducts global redundancy-aware temporal segmentation followed by spatiotemporal merging to reduce the visual token sequence prior to LLM forwarding. Inside the LLM, it further applies token similarity-based merging across layers, so that the two stages jointly remove both video-level and representation-level redundancy in a coordinated manner.

CDPruner Zhang et al. (2026) is a query-driven visual token pruning method for MLLMs. It defines an instruction-conditioned similarity between visual tokens, which measures redundancy with respect to the user instruction rather than purely visual appearance. Token selection is then formulated as a determinantal point process over this instruction-aware kernel, and CDPruner greedily retains a subset of tokens that maximizes the conditional diversity under the given query, yielding compact yet instruction-faithful visual inputs.

QuoTA[‡] Luo et al. (2026) is a query-oriented visual token assignment method for long video understanding in LVLMs. Instead of pruning tokens by visual redundancy or attention-based importance, it distributes the visual token budget across frames according to their relevance to the user query. QuoTA uses a chain-of-thought prompt to decouple the query into a list of query-relevant entities, and a lightweight scoring LVLM then rates each frame in parallel against this list. Under a fixed total budget, these frame-wise scores guide a dynamic token assigner that grants relevant frames more tokens and compresses less relevant ones via bilinear interpolation, producing compact and query-faithful visual inputs in a training-free manner.

[‡]We re-ran essentially all compared methods above to obtain their reported results. For a fair comparison under an identical input protocol, we disable QuoTA's Dynamic Frame Sampling module. When all frames prior to sampling are considered, Table 5 shows that combining our method with MDP3 Sun et al. (2025) attains 58.9 on LongVideoBench, surpassing the 57.4 reported in the original QuoTA paper.

Table 7: Performance comparison of LLaVA-OV-7B with different methods across multiple benchmarks under different random seeds. All methods use the default 32-frame input setting with a retention ratio of 25%.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
Seed=42									
LLaVA-OV-7B	58.4	56.4	63.0	70.1	56.6	48.8	58.5	59.1	100.0
HoliTom[NIPS25]	**58.2**	56.5	62.5	69.6	55.4	46.8	57.3	58.6	99.2
CDPruner[NIPS25]	56.6	54.9	61.2	69.8	54.7	48.6	57.7	57.6	97.5
QuoTA[AAAI26]	57.2	56.8	61.6	68.2	56.4	46.6	57.1	58.2	98.4
CRISP[ICME26]	57.2	56.3	62.9	69.7	54.8	47.9	57.4	58.5	98.9
FastVID[NIPS25]	56.9	56.3	61.0	69.3	55.4	**49.2**	58.0	58.1	98.3
VidCom^2[EMNLP25]	57.1	55.0	62.5	69.4	**56.7**	49.1	**58.4**	58.3	98.6
FlashVID[ICLR26]	57.7	56.6	**63.3**	69.1	54.5	47.7	57.1	58.7	99.3
EviQ(ours)	57.7	**57.4**	62.9	**69.9**	54.5	47.9	57.4	**58.9**	**99.6**
Seed=3407									
LLaVA-OV-7B	58.3	56.4	63.0	70.0	56.6	48.8	58.4	59.0	100.0
HoliTom[NIPS25]	**58.2**	56.7	62.7	69.3	55.8	46.7	57.3	58.7	99.5
CDPruner[NIPS25]	56.6	54.9	61.2	**69.8**	54.7	48.6	57.7	57.6	97.6
QuoTA[AAAI26]	57.2	56.8	61.6	68.2	56.4	46.6	57.1	58.2	98.6
CRISP[ICME26]	57.2	56.3	63.0	69.7	54.8	48.0	57.4	58.5	99.1
FastVID[NIPS25]	56.5	56.1	60.8	69.4	**58.1**	47.2	58.3	57.9	98.1
VidCom^2[EMNLP25]	57.1	55.0	62.5	69.4	56.7	**49.1**	**58.4**	58.3	98.7
FlashVID[ICLR26]	57.7	56.8	**63.5**	69.3	54.8	47.8	57.3	58.8	99.7
EviQ(ours)	57.7	**57.4**	63.0	69.7	55.1	48.2	57.6	**58.9**	**99.8**

F More Experimental Results

F.1 Additional Experiments on LLaVA-OneVision

Comparisons under Different Random Seeds. All experiments in the main text set the Python random seed to *0*, the NumPy seed to *1234*, and the PyTorch manual seed to *1234*. It is worth noting that our evaluation pipeline is fully deterministic given a fixed set of seeds, so repeated runs under the same seed yield identical results with zero variance. We therefore do not report error bars or standard deviations, and instead assess robustness by comparing all methods across multiple distinct seeds under matched settings. To ensure the observed gains of **EviQ** are not due to a particular random initialization, we evaluate all methods under multiple seeds and compare them within each identical random setting. This controls stochastic variation and allows us to assess whether **EviQ** consistently outperforms baselines across matched conditions rather than relying on absolute score differences across seeds. Table 7 reports the performance of LLaVA-OV-7B with different compression methods across multiple benchmarks under two additional random seeds, namely *42* and *3407* Picard (2021). Together with the results under the default seed in Table 1, **EviQ** consistently achieves the best overall performance among the compared methods while retaining at least 99.6% of the accuracy of the uncompressed model using only 25% of the visual tokens. In particular, **EviQ** obtains an average score of 58.9 under both additional seeds, demonstrating that its performance remains stable despite variations in random sampling. Moreover, the consistent improvements under matched random settings indicate that the advantage of **EviQ** is not attributable to a favorable seed or incidental evaluation fluctuation. Instead, the results confirm that its query-aware budget allocation and token selection strategies provide robust and reproducible gains across different benchmarks and random configurations. Table 8 presents analogous multi-seed experiments on another VideoLLM, yielding consistent observations and further confirming the robustness of **EviQ** across model architectures, and we leave out the repetitive discussion for brevity.

F.2 Additional Experiments on LLaVA-Video

Effect of the Number of Input Frames. We further compare the results under different numbers of input frames to analyze the robustness of each pruning method when the visual token budget increases. As shown in Tables 2 and 9, increasing the input from 32 frames to 64 frames improves the full LLaVA-Video-7B model from 60.5 to 62.7 average score, indicating that denser frame sampling provides more useful visual evidence for video understanding. **EviQ** also benefits from the increased number of frames, improving from 59.0 to 60.7 average score. Meanwhile, **EviQ** consistently achieves the best average performance among all pruning methods under both settings, with relative accuracies of 97.6% and 96.8% for 32-frame and 64-frame inputs, respectively. This demonstrates that **EviQ** can effectively scale to larger visual token budgets while preserving most of the

Table 8: Performance comparison of LLaVA-Video-7B with different methods across multiple benchmarks under different random seeds. All methods use the default 64-frame input setting with a retention ratio of 25%.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
Seed=42									
LLaVA-Video-7B	60.1	58.8	67.6	77.3	62.3	53.6	64.4	62.7	100
HoliTom[NIPS25]	**58.0**	57.1	63.7	74.1	60.8	51.6	62.1	60.2	96.1
CDPruner[NIPS25]	56.2	56.1	60.7	72.9	60.1	50.7	61.2	58.6	93.4
QuoTA[AAAI26]	57.1	55.5	63.0	73.1	60.4	51.4	61.7	59.3	94.6
CRISP[ICME26]	56.5	**57.5**	58.5	73.8	60.0	52.3	62.0	58.6	93.5
FastVID[NIPS25]	57.9	56.8	59.7	**75.0**	61.0	**52.7**	**62.9**	59.3	94.6
VidCom^2[EMNLP25]	57.0	56.7	58.9	73.6	**61.7**	50.6	61.9	58.6	93.5
FlashVID[ICLR26]	57.9	56.7	62.2	**75.0**	61.3	51.8	62.7	59.9	95.5
EviQ(ours)	57.9	**57.5**	**63.8**	74.6	60.8	51.1	62.2	**60.4**	**96.3**
Seed=3407									
LLaVA-Video-7B	60.3	58.8	67.6	77.3	62.3	53.6	64.4	62.8	100
HoliTom[NIPS25]	**58.0**	57.1	63.7	73.9	60.7	51.7	62.1	60.2	95.9
CDPruner[NIPS25]	56.2	56.1	60.7	72.9	60.1	50.7	61.2	58.6	93.2
QuoTA[AAAI26]	57.1	55.5	63.0	73.1	60.4	51.4	61.7	59.3	94.5
CRISP[ICME26]	56.5	**57.5**	58.8	73.9	60.0	52.1	62.0	58.7	93.5
FastVID[NIPS25]	57.9	56.3	60.4	74.6	61.2	**52.9**	**62.9**	59.4	94.5
VidCom^2[EMNLP25]	57.0	56.7	58.9	73.6	**61.7**	50.6	61.9	58.6	93.4
FlashVID[ICLR26]	57.9	56.8	62.2	**74.9**	61.2	51.8	62.6	59.9	95.3
EviQ(ours)	57.9	**57.5**	**63.8**	74.6	60.8	51.1	62.2	**60.4**	**96.1**

Table 9: Performance comparison of LLaVA-Video-7B with different methods across multiple benchmarks. All methods use the default **32-frame** input setting with a retention ratio of 25%.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
LLaVA-Video-7B	60.2	57.9	61.1	76.3	59.3	52.2	62.6	60.5	100.0
HoliTom[NIPS25]	**58.0**	55.1	**62.0**	72.9	57.9	49.8	60.2	58.8	97.3
CDPruner[NIPS25]	56.3	55.0	57.2	73.1	58.2	50.8	**60.7**	57.3	94.8
FastVID[NIPS25]	57.9	55.4	57.9	73.1	**58.4**	50.7	**60.7**	58.0	95.9
VidCom^2[EMNLP25]	57.0	52.8	54.8	71.4	56.9	**51.0**	59.8	56.1	92.8
FlashVID[ICLR26]	57.9	55.5	59.6	**73.7**	58.0	50.2	60.6	58.4	96.6
EviQ(ours)	57.9	**56.3**	61.0	72.8	58.3	**51.0**	**60.7**	**59.0**	**97.6**

full-model performance. Compared with other baselines, **EviQ** shows a more stable performance-efficiency trade-off: although some methods perform well on specific benchmarks or VideoMME subsets, their overall performance is less consistent. In contrast, **EviQ** maintains competitive results across MVBench, LongVideoBench, MLVU, and VideoMME, suggesting that query-aware frame-level budget allocation and intra-frame query-coverage token selection are effective for handling both sparse and dense frame inputs.

More Analysis on Efficiency. We further evaluate different training-free token pruning methods on LLaVA-Video-7B under the 64-frame input setting, as shown in Table 10. Compared with the LLaVA-OV setting, this scenario involves longer visual inputs and therefore imposes a more challenging accuracy-efficiency trade-off. Existing methods exhibit different limitations: lightweight methods such as FastVID and VidCom^2 achieve the highest speedups of 1.71× and 1.72×, respectively, but suffer from larger performance drops, while accuracy-preserving methods such as HoliTom and FlashVID retain relatively better performance at the cost of higher additional overhead and lower acceleration. CDPruner introduces substantial overhead, resulting in almost no speedup. In contrast, **EviQ** achieves a strong 1.66× speedup while maintaining the best performance among all pruning methods, with only a 2.0-point drop from the original model. This demonstrates that **EviQ** provides a more favorable accuracy-efficiency trade-off on LLaVA-Video-7B. Moreover, **EviQ** reduces the average GPU memory usage by 54.7%, showing its practical benefit for long-video inference.

F.3 Additional Experiments on Qwen3-VL

Robustness in Compression Ratio. Table 11 reports the results of **EviQ** and seven representative video token pruning baselines on Qwen3-VL-8B-Instruct under a 35% retention ratio. **EviQ** achieves the highest average relative score of 98.6% and ties for the best absolute average score of 63.2 with CDPruner, while clearly outperforming the remaining baselines. In particular, it delivers

Table 10: Efficiency comparison of LLaVA-Video-7B with different methods across a total of four benchmarks. All methods use the default 64-frame input setting with a retention ratio of 25%.

Method	Encode Latency(ms)	Prefill Latency(ms)	Overhead Latency(ms)	Total Latency(ms)	GPU Average Memory(MB)	Performance
LLaVA-Video-7B	174.7	336.1	0.0	545.8	56199	62.7
HoliTom[NIPS25]	246.3	104.3	71.7	384.3(×1.42)	25449(↓54.7%)	59.8(↓2.9)
CDPruner[NIPS25]	414.1	103.7	239.3	552.8(×0.99)	25451(↓54.7%)	58.6(↓4.1)
QuoTA[AAAI26]	4668.7	103.6	4490.1	4806.7(×0.11)	33725(↓40.0%)	59.3(↓3.4)
CRISP[ICME26]	211.5	104.8	48.5	367.8(×1.48)	29517(↓47.5%)	58.7(↓4.0)
FastVID[NIPS25]	181.9	102.7	7.2	319.5(×**1.71**)	25449(↓54.7%)	59.4(↓3.3)
VidCom^2[EMNLP25]	179.0	104.3	4.4	317.5(×**1.72**)	25449(↓54.7%)	58.5(↓4.2)
FlashVID[ICLR26]	226.7	105.3	52.0	366.7(×1.49)	25451(↓54.7%)	59.9(↓2.8)
EviQ(ours)	190.2	103.9	15.6	328.6(×**1.66**)	25449(↓54.7%)	60.7(↓**2.0**)

Table 11: Performance comparison of Qwen3-VL-8B-Instruct with different methods across multiple benchmarks. All methods use the default 32-frame input setting.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
Qwen3-VL-8B-Instruct	68.6	60.2	63.4	76.0	60.0	56.9	64.3	64.1	100.0
Retention Ratio=35%									
HoliTom[NIPS25]	65.0	57.7	61.9	71.9	57.5	53.9	61.1	61.4	95.8
CDPruner[NIPS25]	68.1	58.3	**62.6**	74.8	59.2	**56.8**	**63.6**	**63.2**	98.5
FastVID[NIPS25]	67.5	58.6	62.2	73.1	58.7	53.5	61.7	62.5	97.5
VidCom^2[EMNLP25]	68.1	58.5	61.9	73.6	**59.5**	55.0	62.7	62.8	97.9
MMG-Vid[AAAI26]	68.2	58.9	61.3	72.3	57.7	54.4	61.5	62.5	97.4
FlashVID[ICLR26]	67.9	59.1	62.3	**75.2**	59.2	53.2	62.5	63.0	98.2
ForestPrune[CVPR26]	62.2	58.9	60.6	73.6	57.0	55.1	61.9	60.9	95.0
EviQ(ours)	**68.4**	**60.1**	62.5	72.7	58.3	54.6	61.8	**63.2**	**98.6**

the best results on MVBench with a score of 68.4 and on LongVideoBench with a score of 60.1, recovering the performance of the uncompressed Qwen3-VL almost exactly on the latter. These observations are consistent with Table 3 in the main text, which further reports results at the more aggressive 25% and 15% retention ratios and shows that **EviQ** degrades more gracefully as the retention ratio decreases. Taken together, the results confirm that **EviQ** generalizes well to a stronger VideoLLM backbone and remains effective across a wide range of compression levels.

Scalability to Larger VideoLLMs. To further examine whether **EviQ** remains effective beyond 7B/8B-scale models, we conduct additional experiments on the larger Qwen3-VL-32B-Instruct model, as shown in Table 12. Under the default 32-frame input setting with a retention ratio of 25%, **EviQ** achieves the highest average score among all compared pruning methods, reaching 66.4 and retaining 98.0% of the full-model performance. Compared with strong baselines such as FlashVID and VidCom^2, **EviQ** improves the average score by 0.4 and 0.8 points, respectively. In particular, **EviQ** obtains the best results among pruning methods on MVBench, LongVideoBench, and MLVU, and even slightly surpasses the full model on LongVideoBench. Although **EviQ** is not always the best on every VideoMME subset, it delivers the strongest overall performance across benchmarks. These results demonstrate that **EviQ** scales well to larger VideoLLMs and remains effective in preserving task-relevant visual information under substantial token compression.

F.4 Detailed Ablation Study Results

Effect of EviQ components. Table 13 provides further insight into the design choices of **EviQ**. **(i) Embedding norm serves as an effective informativeness proxy.** Across benchmarks, the ℓ_2-norm consistently outperforms channel variance and activation-distribution entropy, confirming that embedding norm is an effective proxy for visual informativeness. **(ii) Marginal query coverage complements context preservation.** QCE alone performs substantially worse, suggesting that prioritizing query matches can discard the supporting context and broader scene information needed for reasoning. Combining QCE with CPE improves average performance across all budgeting settings: the visual informativeness prior retains supporting context, while query coverage compensates for CPE's lack of query awareness. Replacing only the QCE term with a sum of individual token relevance scores while retaining CPE lowers the average score from 59.0 to 57.0, highlighting the value of marginal query coverage over independent relevance scoring. Such a formulation avoids repeatedly rewarding already well-matched query tokens and instead promotes complementary evidence.

Table 12: Performance comparison of Qwen3-VL-32B-Instruct with different methods across multiple benchmarks. All methods use the default 32-frame input setting with a retention ratio of 25%.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
Qwen3-VL-32B-Instruct	73.2	62.3	66.1	78.3	67.1	62.3	69.3	67.7	100.0
HoliTom[NIPS25]	62.3	60.2	63.2	74.2	61.2	59.1	64.9	62.6	92.5
CDPruner[NIPS25]	69.9	60.9	63.8	74.9	62.1	61.0	66.0	65.2	96.2
FastVID[NIPS25]	71.0	60.9	64.5	75.1	62.6	57.7	65.1	65.4	96.6
VidCom^2[EMNLP25]	70.2	60.4	64.9	75.6	**64.3**	61.2	67.0	65.6	96.8
MMG-Vid[AAAI26]	64.4	59.1	63.1	73.2	62.7	59.9	65.3	63.0	93.0
FlashVID[ICLR26]	70.7	61.4	64.5	**76.7**	63.2	**61.3**	**67.1**	66.0	97.5
EviQ(ours)	**71.6**	**62.8**	**65.1**	75.6	63.9	58.4	66.0	**66.4**	**98.0**

Table 13: Ablation study of **EviQ** on LLaVA-OV-7B at 25% retention. We group ablations by module removal, uniform per-frame budgeting, and design variants. For QCBR, subscripts denote frame-level query relevance pooling: mean averaging, max pooling, and log-sum-exp pooling. For CPE, ℓ_2, var, and ent denote token scoring by feature norm, channel variance, and activation-distribution entropy. Uniform assigns the same retention ratio to every frame. RelSort replaces QCE with summed token relevance scores while retaining CPE, ranking tokens by combined informativeness and relevance. [noun-only] uses noun-filtered query tokens.

Method	MVBench	LongVideoBench	MLVU	VideoMME	Average	
					Score	Rel.(%)
Vanilla	58.3	56.4	63.0	58.4	59.0	100.0
Module ablation						
CPE_{ℓ_2}	57.5	57.1	62.6	56.1	58.3	98.9
QCE	51.0	49.6	55.2	52.7	52.1	88.3
CPE_{ℓ_2} + QCE	57.5	57.2	62.7	56.6	58.5	99.2
$\text{QCBR}_{\text{mean}}$ + CPE_{ℓ_2}	57.7	57.1	62.4	56.7	58.5	99.1
$\text{QCBR}_{\text{mean}}$ + QCE	54.5	55.0	61.2	57.0	56.9	96.4
Uniform per-frame budget						
Uniform + CPE_{ℓ_2}	57.5	57.1	62.2	56.7	58.4	98.9
Uniform + QCE	53.9	54.9	61.1	56.9	56.7	96.1
Uniform + CPE_{ℓ_2} + QCE	57.6	57.2	62.6	57.5	58.7	99.5
QCBR, CPE, and QCE variants						
$\text{QCBR}_{\text{mean}}$ + CPE_{var} + QCE	57.2	56.6	60.2	57.3	57.8	98.0
$\text{QCBR}_{\text{mean}}$ + CPE_{ent} + QCE	55.0	53.6	60.2	56.1	56.2	95.3
QCBR_{max} + CPE_{ℓ_2} + QCE	57.3	55.9	61.2	58.0	58.1	98.4
QCBR_{LSE} + CPE_{ℓ_2} + QCE	57.5	55.7	61.0	57.7	58.0	98.2
$\text{QCBR}_{\text{mean}}$ + CPE_{ℓ_2} + RelSort	56.5	54.1	60.6	56.6	57.0	96.5
$\text{QCBR}_{\text{mean}}$ + CDPruner[NIPS25]	57.2	55.8	61.2	57.8	58.0	98.1
$\text{QCBR}_{\text{mean}}$ + CRISP[ICME26]	57.4	56.6	62.9	57.4	58.6	99.1
$\text{QCBR}_{\text{mean}}$ + CPE_{ℓ_2} + $\text{QCE}_{[\text{noun-only}]}$	57.1	56.6	62.6	58.2	58.6	99.4
EviQ ($\text{QCBR}_{\text{mean}}$ + CPE_{ℓ_2} + QCE)	57.7	57.4	63.0	57.8	59.0	99.9

(iii) Query-conditioned frame budgets improve both efficiency and accuracy. QCBR decomposes global greedy selection into local greedy selection within each frame, reducing selection time by 61.6% while improving average performance via explicit frame budgets that allocate more capacity to query-relevant frames. Its advantage over uniform budgeting further confirms the benefit of adapting frame budgets to query relevance. Among the aggregation variants, mean pooling attains higher average accuracy than max and log-sum-exp pooling, motivating its use as the default frame-relevance estimator. Taken together, these results justify combining a visual informativeness prior, marginal query coverage, and adaptive frame budgets for efficient token pruning.

Effect of query construction on evidence selection. For relevance estimation, we use the LLM input embeddings of all text tokens following the image token in the formatted prompt, including the question, candidate options, response instructions, and any subsequent conversation-template tokens. No additional linguistic filtering or word-level pooling is applied, so relevance is computed over individual tokenizer tokens. To assess the effect of noun filtering, we construct a noun-only query using the noun-extraction procedure adopted by CRISP Li et al. (2026). As shown in Table 13, this variant lowers the average score from 59.0 to 58.6, with decreases of 0.6, 0.8, and 0.4 points on MVBench, LongVideoBench, and MLVU, respectively. These results suggest that noun filtering may discard useful cues for video evidence selection, as non-noun words can express actions, attributes, and temporal relations beyond object identity. Conversely, VideoMME improves from 57.8 to 58.2, suggesting that filtering may benefit some settings by reducing irrelevant textual

Table 14: Performance comparison of hyperparameters under different retention ratios across MVBench, LongVideoBench, MLVU, and VideoMME, including concentration coefficient τ in Eq.4 and control parameter α in Eq.5.

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Short	Medium	Long	Overall	Score	Rel.(%)
Qwen3-VL-8B-Instruct	68.6	60.2	63.4	76.0	60.0	56.9	64.3	64.1	100
Retention Ratio=35%									
τ=0.0,α=1.0	68.2	60.1	62.1	72.1	58.0	53.0	61.0	62.9	98.0
τ=1.0,α=1.0	68.2	60.1	62.1	72.4	57.7	53.6	61.2	62.9	98.1
τ=1.5,α=1.0	68.2	60.0	62.0	71.7	57.3	53.1	60.7	62.7	97.9
τ=0.5,α=0.0	68.5	59.7	62.1	72.0	58.1	54.2	61.4	62.9	98.1
τ=0.5,α=0.5	68.5	60.1	62.3	72.0	58.3	53.6	61.3	63.1	98.3
τ=0.5,α=1.5	68.4	60.4	62.0	72.0	57.9	54.3	61.4	63.1	98.3
τ=0.5,α=1.0	68.4	60.1	62.5	72.7	58.3	54.6	61.8	63.2	98.6
Retention Ratio=25%									
τ=0.0,α=1.0	67.8	59.8	61.6	70.9	56.9	53.8	60.5	62.4	97.3
τ=1.0,α=1.0	67.6	59.8	61.4	71.6	57.1	53.2	60.6	62.4	97.4
τ=1.5,α=1.0	67.8	59.2	61.7	71.8	56.7	53.7	60.7	62.4	97.4
τ=0.5,α=0.0	67.6	60.1	61.4	70.4	57.4	54.0	60.6	62.4	97.3
τ=0.5,α=0.5	67.5	59.9	61.3	71.6	56.4	53.6	60.5	62.3	97.2
τ=0.5,α=1.5	67.7	59.6	62.0	71.4	56.5	53.8	60.6	62.4	97.3
τ=0.5,α=1.0	67.9	59.8	61.7	71.9	57.6	53.1	60.9	62.6	97.7
Retention Ratio=15%									
τ=0.0,α=1.0	65.1	58.7	60.5	68.3	55.2	52.9	58.8	60.8	94.8
τ=1.0,α=1.0	66.2	59.1	60.4	68.3	54.9	53.1	58.8	61.1	95.4
τ=1.5,α=1.0	66.3	58.6	60.3	68.1	55.6	52.6	58.7	61.0	95.1
τ=0.5,α=0.0	66.1	58.4	60.7	67.3	56.1	53.1	58.9	61.0	95.2
τ=0.5,α=0.5	66.1	58.9	60.6	68.4	56.3	53.9	59.6	61.3	95.6
τ=0.5,α=1.5	66.4	58.7	60.4	68.4	56.2	52.8	59.1	61.2	95.4
τ=0.5,α=1.0	66.2	59.2	60.8	69.1	56.3	54.1	59.9	61.5	96.0

cues. Overall, retaining the text without noun filtering achieves better average performance while avoiding an additional noun-extraction step, supporting our default query construction.

Effect of the concentration coefficient τ and the balance parameter α. Figure 4 in the main text illustrates the robustness of **EviQ** to the two hyperparameters, and Table 14 provides detailed results on Qwen3-VL-8B-Instruct under three retention ratios. The concentration coefficient τ controls the sharpness of frame-level budget routing, with τ=0 degenerating to uniform allocation and disabling frame-level query-conditioned routing. Across all retention ratios, varying τ induces only marginal fluctuations on MVBench and LongVideoBench. The balance parameter α controls the trade-off between Context-Preserving Evidence (CPE) and Query-Complete Evidence (QCE), with α=0 reducing intra-frame selection to CPE-only token selection. Varying α likewise leads to limited changes on MLVU and VideoMME. The sensitivity to both hyperparameters grows slightly as the retention ratio decreases, but overall performance remains within a narrow band, confirming that **EviQ** is robust to τ and α. Among all configurations, τ=0.5 and α=1.0 achieve the best average performance and are adopted as the default setting.

G More Visualizations

G.1 Visualizations of EviQ

Figure 6 illustrates the query-aware token pruning process of **EviQ** across diverse video question-answering examples. The first row shows the uniformly sampled input frames; the second row presents the query-video relevance estimated from the aligned visual and query embeddings; and the third row shows the final retained visual tokens, with orange boxes indicating the key frames required to answer each question. Across temporal-ordering, action-recognition, event-understanding, and subtitle-dependent examples, our method consistently assigns higher relevance scores to frames containing query-critical evidence and preserves denser token regions within these frames. These visualizations demonstrate that frame-level budget allocation effectively prioritizes query-relevant frames, while the intra-frame selection strategy preserves informative regions with comprehensive query coverage.

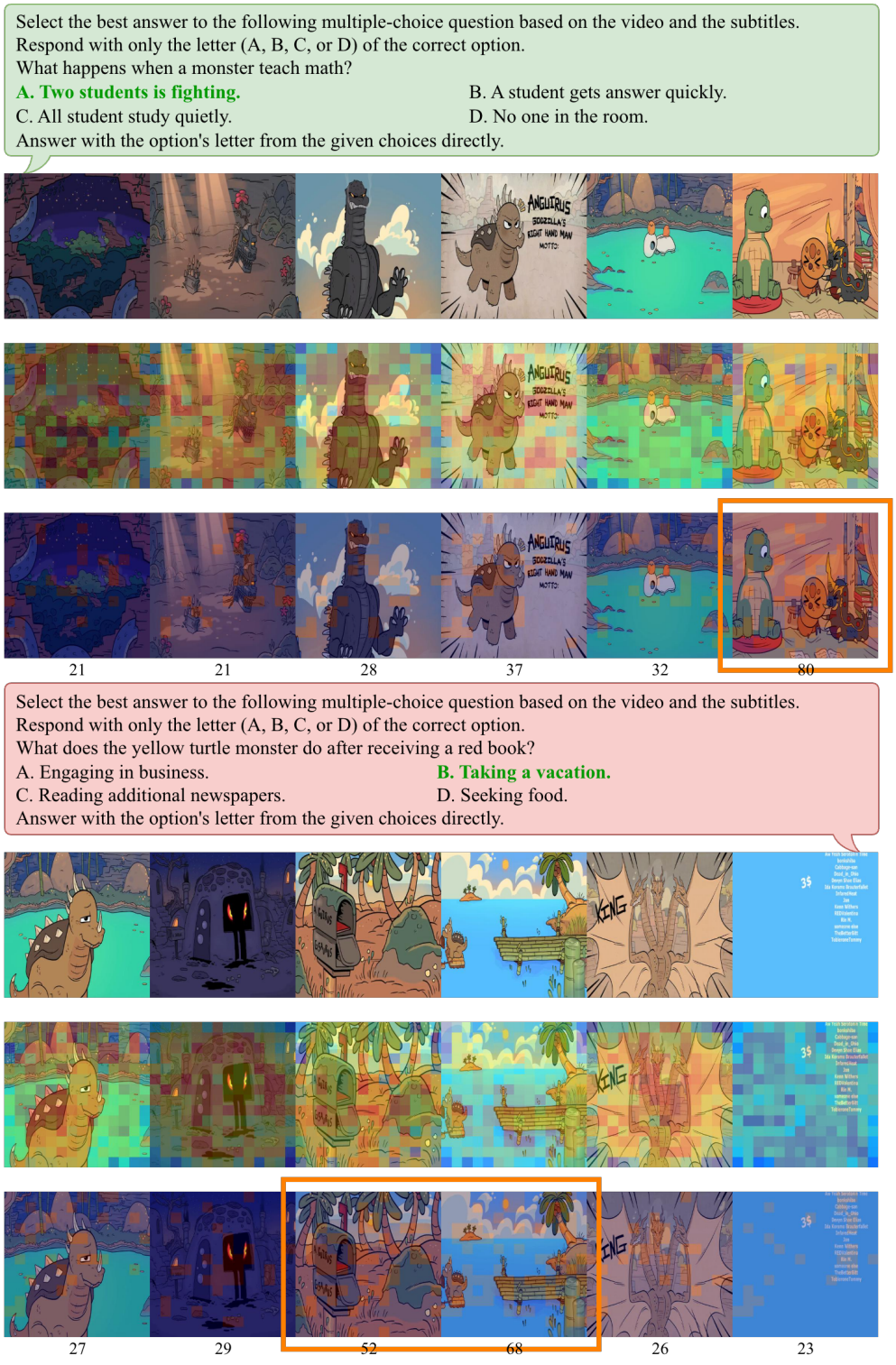
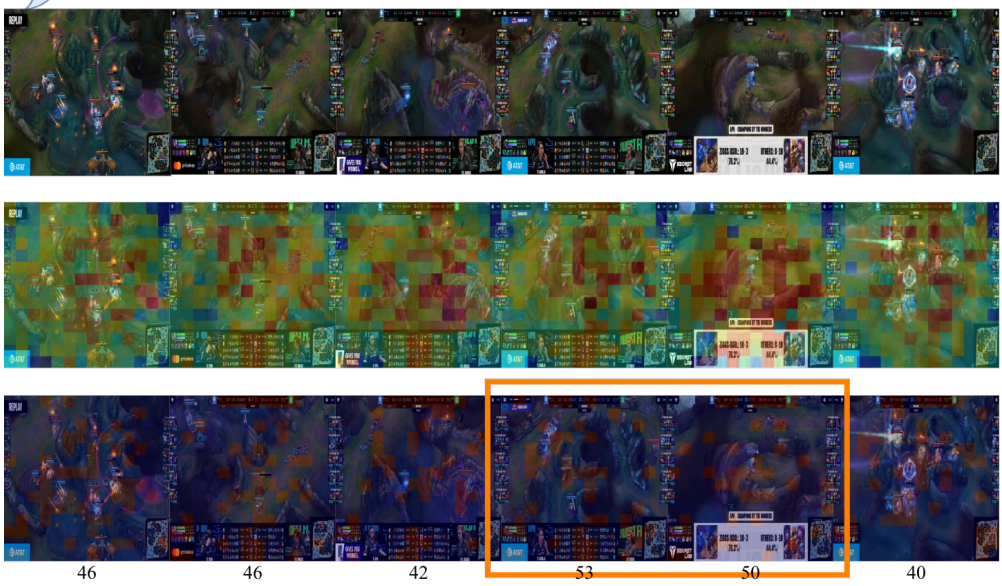


Figure 6: Visualizations of **EviQ**. The first row shows the original images. The second row visualizes query–video relevance through embedding similarity and the color gradient indicates the degree of relevance, where cooler blue tones represent lower correlation and warmer red tones denote higher relevance. The third row shows the final token selection results based on the relevance above and the red blocks indicate selected tokens. The sentences in green are the selected options that are also correct. The numbers below indicate the number of selected tokens.

Select the best answer to the following multiple-choice question based on the video and the subtitles.
Respond with only the letter (A, B, C, or D) of the correct option.
Who will lead their team to victory in this game?
A. Inspired by team FLY.
B. Massu from team FLY.
C. Yeon from team TL.
D. APA from team TL.
Answer with the option's letter from the given choices directly.



Select the best answer to the following multiple-choice question based on the video and the subtitles.
Respond with only the letter (A, B, C, or D) of the correct option.
What happens when it is at ban/pick time?
A. Team TL is banned.
B. Team FLY gains a more ban.
C. Team FLY doesn't ban a hero at its turn.
D. Team TL misses a ban.
Answer with the option's letter from the given choices directly.

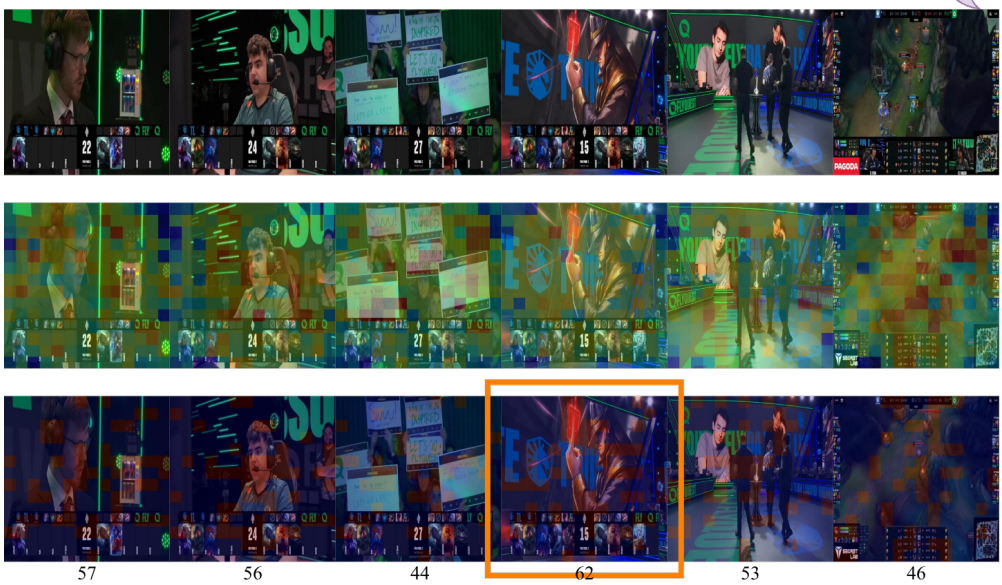
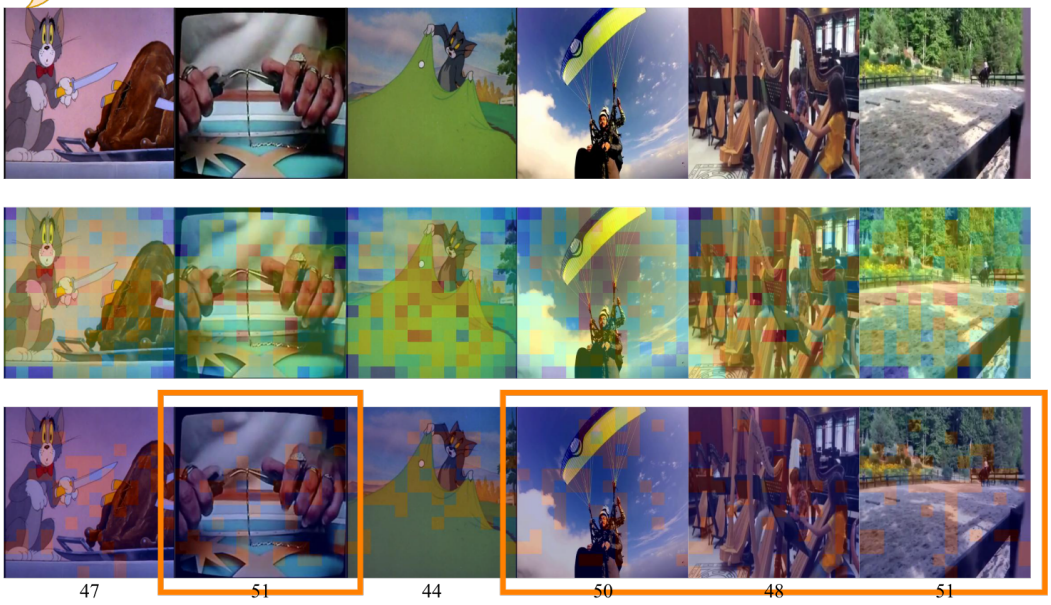


Figure 6: (continued)

Please identify the option that corresponds to the order of events as they occur in the video.
(A) paragliding --> playing harp --> riding mule --> making jewelry
(B) making jewelry --> paragliding --> playing harp --> riding mule
(C) paragliding --> riding mule --> playing harp --> making jewelry
(D) riding mule --> making jewelry --> paragliding --> playing harp
Only give the best option.
Best option: (



What are the cartoon starfish and cartoon sponge doing while they are talking?
(A) Playing with toys
(B) Drinking water
(C) Eating snacks
(D) Riding a cartoon seahorse seat
Only give the best option.
Best option: (



Figure 6: (continued)

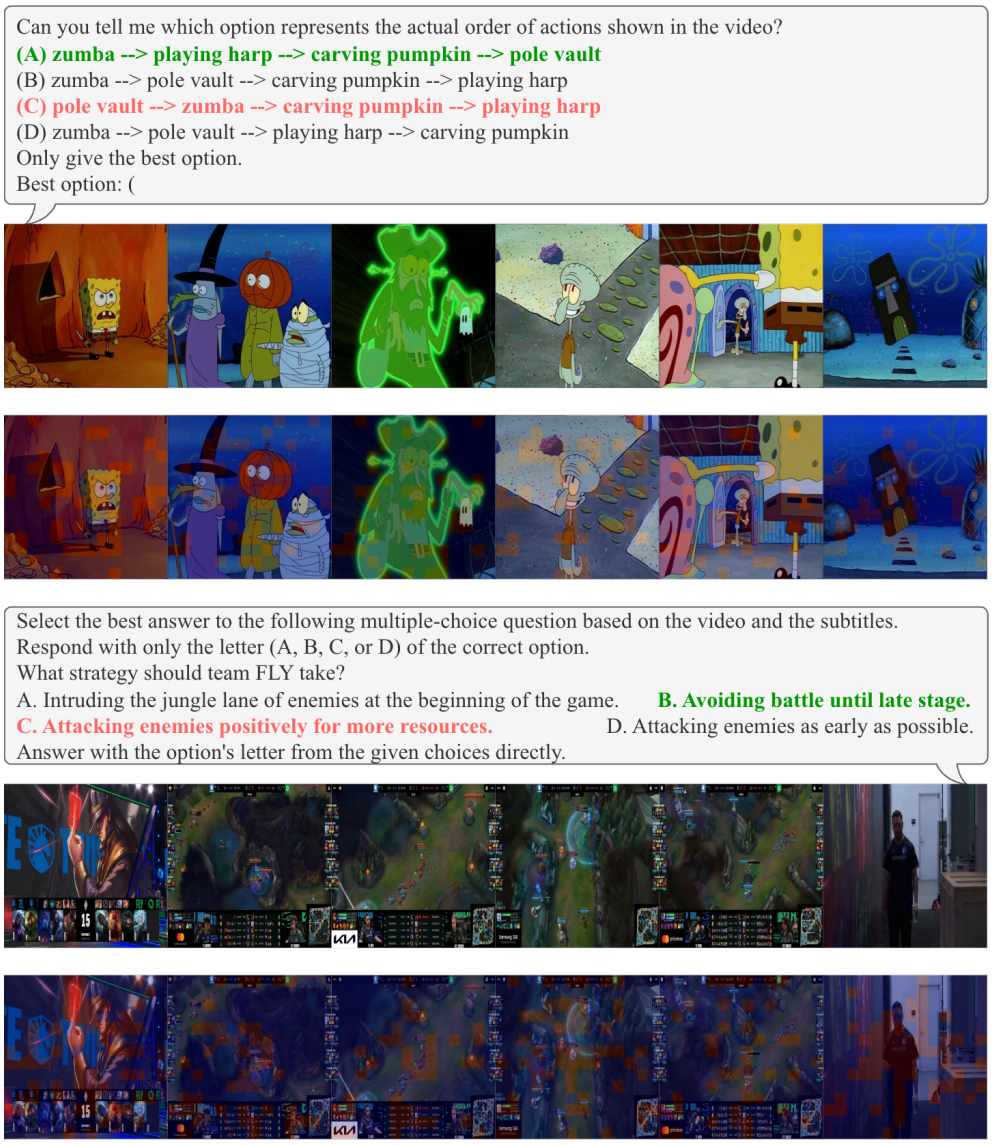


Figure 7: Visualizations of failure cases in **EviQ**. For each sample, the top row shows the original frames and the bottom row the final token selection, where red blocks mark the selected tokens. Red sentences denote incorrectly selected options and green ones the correct answers.

G.2 Visualizations of Failure Cases

Figure 7 illustrates two representative failure modes of **EviQ**. First, for extremely long videos, uniform temporal downsampling may discard query-relevant events before token pruning is even applied. In the first example, none of the sampled frames contain the actions mentioned in the query, making it impossible for **EviQ** to recover the missing evidence regardless of how accurately it allocates or selects visual tokens. This limitation can be mitigated by adopting more advanced frame sampling strategies Tang et al. (2025); Sun et al. (2025). As shown in Section 4.2, **EviQ** is orthogonal to such methods and generalizes well across different sampling schemes. Second, some questions require complex temporal or strategic reasoning that goes beyond merely identifying query-relevant evidence. As shown in the second example, **EviQ** correctly retains visual tokens associated with the game state, yet the backbone model still predicts the wrong strategy. Notably, the uncompressed model also predicts incorrectly, indicating that the final error stems primarily from the VideoLLMs' inherent reasoning capability rather than the token selection process.