SEEK: PRESERVING SPARSE EVIDENCE ACROSS SEMANTIC EVENTS FOR LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video understanding challenges Multimodal Large Language Models (MLLMs) to identify sparse yet decisive evidence from temporal contexts under limited visual budgets. Keyframe selection provides an efficient way to reduce the visual input while preserving task-relevant information. However, existing methods may miss sparse query-relevant evidence, over-select redundant high-relevance frames, or struggle to handle heterogeneous events with different durations, relevance distributions, and evidence densities, leading to less reliable event-level keyframe allocation. We propose **SEEK**, an event-centric keyframe selection framework for preserving sparse evidence across semantic events. **SEEK** first partitions a video into variable-length events based on local and contextual visual changes. It then evaluates each event by jointly considering peak relevance, top-k mean relevance, and average relevance, and employs the resulting utility to guide event-wise keyframe allocation. Finally, relevance entropy is used within each event to adapt the candidate pool and balance query relevance, visual diversity, and temporal diversity. Experiments on LongVideoBench, Video-MME, and MLVU with LLaVA-Video-7B, InternVL3.5-8B, and Qwen3-VL-8B show that **SEEK** consistently outperforms representative frame selection methods, with improvements of 2.6%–6.7% over uniform temporal sampling.

1 INTRODUCTION

Multimodal large language models (MLLMs) have recently demonstrated strong capabilities in video understanding and long-video question answering. However, processing long videos remains challenging because the number of visual tokens rapidly grows with video duration, while the context capacity and computational budget of current MLLMs remain limited (Shen et al., 2025; Tang et al., 2025). This challenge is especially pronounced on recent long-video benchmarks such as LongVideoBench (Wu et al., 2024), Video-MME (Fu et al., 2025a), and MLVU (Zhou et al., 2025), where models are required to retrieve and reason over sparse evidence distributed across temporal contexts. Consequently, selecting a compact subset of frames that preserves task-relevant evidence has become a critical component for efficient long-video reasoning (Tang et al., 2025; Huang et al., 2025; Peng et al., 2026).

Despite recent progress, existing keyframe selection strategies still face several limitations, as illustrated in Fig. 1. First, *query-agnostic sampling*, including uniform temporal sampling (Cheng et al., 2024) and visual-content-based frame selection (Li et al., 2026) that relies solely on visual information, does not explicitly account for the question when deciding which frames to retain. Although such strategies provide broad temporal coverage, they may miss sparse query-relevant evidence. Second, *query-aware selection* improves task relevance by explicitly conditioning frame selection on the question. A common strategy is to rank individual frames according to query–frame similarity and select frames with higher relevance (Zhang et al., 2025a). However, high-relevance frames are often temporally adjacent and visually similar, resulting in substantial redundancy in the selected set. Recent methods extend frame selection beyond frame-level relevance by incorporating temporal diversity and temporal structure. AKS (Tang et al., 2025) balances query relevance with video coverage, while QCA (Peng et al., 2026) partitions the video into fixed temporal segments and performs query-aware keyframe selection within each segment. A remaining challenge is that different events in long videos often exhibit substantially different durations, relevance distributions, and evidence densities. Fixed temporal segments may not align with semantic event boundaries, causing a coher-

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.

ent event to be split across multiple segments or semantically different content to be grouped within the same segment. This mismatch can make event-wise keyframe allocation less reliable.

These observations suggest that, beyond frame relevance and temporal coverage, long-video keyframe selection also requires *event-level sparse evidence preservation*. This is crucial for retaining sparse yet decisive evidence across heterogeneous events while maintaining sufficient temporal coverage. To this end, we propose **SEEK** (**S**parse-**E**vidence-aware **E**vent **K**eyframe Selection), an event-centric framework following an *event-to-evidence* design. First, *Semantic Event Partition* constructs variable-length events based on local and contextual visual changes, allowing the temporal structure to better align with semantic transitions. Second, *Sparse-Evidence Utility* evaluates each event by jointly considering peak relevance, top-k mean relevance, and average relevance, and guides event-wise keyframe allocation. Third, *Adaptive Selection* selects frames within each event by using relevance entropy to adapt the candidate pool and balance query relevance, visual diversity, and temporal diversity.

We evaluate our method on LongVideoBench (Wu et al., 2024), Video-MME (Fu et al., 2025a), and MLVU (Zhou et al., 2025) with three representative MLLM backbones. **SEEK** achieves the best performance across nearly all evaluated model–benchmark combinations, improving over uniform sampling by 2.6%–6.7%. Compared with QCA (Peng et al., 2026), a closely related segment-based query-aware selection method, **SEEK** further achieves gains of 0.8%–2.9% across all nine evaluation settings.

2 Related Work

Query-Aware Keyframe Selection. Query-aware frame selection aims to retain frames that are relevant to the input query. Q-Frame (Zhang et al., 2025a) uses text–image matching and Gumbel-Max sampling together with adaptive resolution allocation, while FRAG (Huang et al., 2025) scores frames by query relevance and selects high-relevance frames for downstream generation. BOLT (Liu et al., 2025) studies query–frame similarity-based sampling and adopts inverse transform sampling to improve frame selection. Retrieval-based methods further introduce query decomposition or temporal context: E-VRAG (Xu et al., 2025) combines query decomposition with distribution-aware retrieval, while OneClip-RAG (Chen et al., 2025) performs query-guided video chunking followed by clip retrieval.

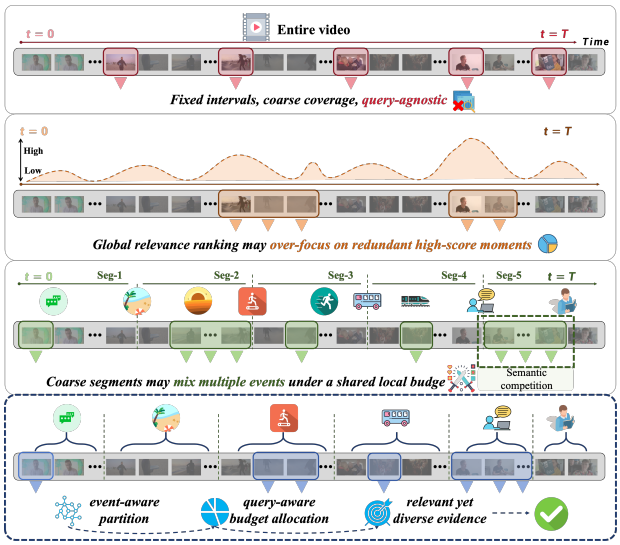

Figure 1: Motivation of SEEK. Existing sampling strategies may miss sparse evidence, select redundant high-relevance frames, or produce unreliable event-wise allocation, whereas SEEK constructs variable-length semantic events for keyframe selection.

Beyond frame relevance, several methods incorporate coverage, diversity, and temporal structure. AKS (Tang et al., 2025) balances query relevance with video coverage, and MDP^3 (Sun et al., 2025) jointly considers relevance, diversity, and temporal order when allocating frames across segments. Nar-KFC (Fang et al., 2026) formulates relevance–diversity selection as integer quadratic programming and supplements selected frames with textual information from non-keyframes. FOCUS (Zhu et al., 2026) explores informative temporal regions through a bandit-based strategy. QCA Peng et al. (2026) allocates frames across fixed temporal segments using query relevance and content variation, followed by diversity-aware selection. WFS-SB (Chen et al., 2026) detects

Algorithm 1 Semantic Event Partition

Require: Visual features $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, boundary percentile P_b, minimum event length $L_{\min}$, maximum event length $L_{\max}$
Ensure: Event set $\{\mathcal{E}_e\}_{e=1}^{M}$

Compute the temporal change signal $\mathbf{B} = \{B_t\}_{t=1}^{N-1}$ using Eq. 2
$\theta_b \leftarrow \text{Percentile}_{P_b}(\mathbf{B})$
$\mathcal{C} \leftarrow \{t \mid B_t \geq \theta_b,\ B_t \geq B_{t-1},\ B_t \geq B_{t+1}\}$
$\mathcal{B} \leftarrow \emptyset$
for t in $\mathcal{C}$ sorted by descending B_t **do**
    if $|t - b| \geq L_{\min},\ \forall b \in \mathcal{B}$ **then**
        $\mathcal{B} \leftarrow \mathcal{B} \cup \{t\}$
    end if
end for
Sort $\mathcal{B}$ in temporal order and append endpoints 0 and N
for each interval $[a, b)$ induced by consecutive boundaries **do**
    while $b - a > L_{\max}$ **do**
        $\Omega \leftarrow \{t \mid a + L_{\min} \leq t \leq \min(a + L_{\max},\ b - L_{\min})\}$
        $t^* \leftarrow \arg\max_{t \in \Omega} B_t$
        Insert t^* into $\mathcal{B}$
        $a \leftarrow t^*$
    end while
end for
Form $\mathcal{E}_e = \{a_{e-1} + 1, \ldots, a_e\}$ from the sorted boundaries $0 = a_0 < \cdots < a_M = N$
return $\{\mathcal{E}_e\}_{e=1}^{M}$

semantic boundaries from query–frame similarity signals, allocates frames across the resulting clips, and applies maximal marginal relevance within each clip.

Learning-based approaches optimize frame selection using supervision or downstream task feedback. MLLM-Frame (Hu et al., 2025) trains a lightweight selector with spatial and temporal supervision, while VideoMiner (Cao et al., 2025), RL-Frame (Qin et al., 2026), and MSJoE (Tan et al., 2026) learn selection policies through reinforcement learning or joint optimization with the MLLM.

Compared with existing methods, our approach focuses on *event-level sparse evidence preservation*. **SEEK** first constructs variable-length semantic events according to local and contextual visual changes, allowing keyframe allocation to better follow the underlying temporal structure of the video. It then jointly considers peak relevance, top-k mean relevance, and average relevance within each event to capture sparse decisive evidence and guide event-wise keyframe allocation. Finally, relevance entropy is used to adapt the candidate pool and balance query relevance, visual diversity, and temporal diversity during within-event selection.

3 METHOD

3.1 OVERVIEW AND PRELIMINARIES

Given a long video V and a query q, our goal is to select a compact set of frames that preserves query-relevant evidence across events. As illustrated in Fig. 2, our framework comprises three stages. *Semantic Event Partition* constructs variable-length temporal units based on visual transitions. *Sparse-Evidence Utility* evaluates each event using multiple statistics of its frame-level relevance and uses the resulting utility to guide event-wise keyframe allocation. *Adaptive Selection* then selects frames within each event by using relevance entropy to adapt the candidate pool and balance query relevance, visual diversity, and temporal diversity.

Specifically, we sample N temporally ordered candidate frames $\{I_i\}_{i=1}^{N}$ at 1 FPS from the video. For each candidate frame, we extract a frame-level visual representation and a query-conditioned image–text matching (ITM) score using a pretrained BLIP model (Li et al., 2022). The visual representation $\mathbf{x}_i \in \mathbb{R}^d$ is used in the *Semantic Event Partition* stage to capture query-agnostic temporal structures by measuring local and contextual visual changes. The ITM score $r_i \in [0, 1]$ provides the frame-level relevance to the query for computing *Sparse-Evidence Utility* and guiding event-wise keyframe allocation. During *Adaptive Selection*, frame-level relevance is used for candidate rank-

ing and relevance-aware selection. We construct a within-event relevance distribution from these scores and compute its normalized entropy, which is then used to adapt the candidate pool and the relevance–diversity balance.

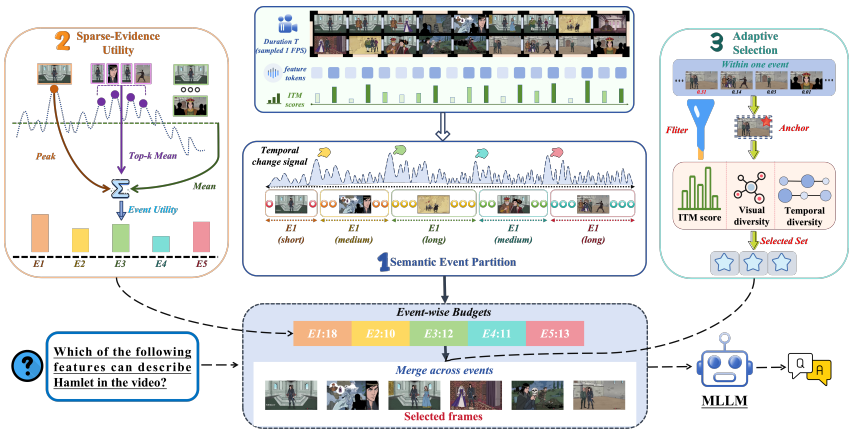

Figure 2: Overview of the proposed framework. *Semantic Event Partition* identifies temporal transitions from local and contextual visual changes to obtain variable-length events. *Sparse-Evidence Utility* combines peak relevance, top-k mean relevance, and mean relevance to characterize sparse evidence within each event and guide event-wise keyframe allocation. *Adaptive Selection* computes relevance entropy from the within-event relevance distribution to adapt the candidate pool and selection weights, then initializes the highest-relevance frame as an anchor and iteratively selects the remaining frames by balancing query relevance, visual diversity, and temporal diversity.

3.2 Semantic Event Partition

To avoid imposing equal-length segments on heterogeneous video content, we construct variable-length temporal partitions based on the temporal change signal derived from visual features. At a boundary after candidate t, the adjacent-frame change is $d_t^{\text{adj}} = 1 - \mathbf{x}_t^\top \mathbf{x}_{t+1}$.

We also compare the visual context on either side of the boundary. Let $\mathcal{W}_t^- = \{\max(1, t-w+1), \dots, t\}$ and $\mathcal{W}_t^+ = \{t+1, \dots, \min(N, t+w)\}$ denote two local windows. Their mean embeddings and contextual change are

$$\boldsymbol{\mu}_t^{\pm} = \frac{1}{|\mathcal{W}_t^{\pm}|} \sum_{i \in \mathcal{W}_t^{\pm}} \mathbf{x}_i, \qquad d_t^{\text{ctx}} = 1 - \frac{(\boldsymbol{\mu}_t^-)^\top \boldsymbol{\mu}_t^+}{\|\boldsymbol{\mu}_t^-\|_2 \|\boldsymbol{\mu}_t^+\|_2}. \tag{1}$$

If either contextual mean has negligible norm, we use d_t^{adj} instead. The two signals capture immediate visual changes and differences between neighboring contexts, respectively. We combine and smooth them to obtain the boundary score

$$B_t = \text{MA}_s\big(\alpha \mathbf{d}^{\text{adj}} + (1-\alpha)\mathbf{d}^{\text{ctx}}\big)_t, \tag{2}$$

where MA_s is a moving average with kernel size s. Based on the above temporal change signal, we identify semantic event boundaries using Algorithm 1. Since this process relies solely on visual features, the event partition can be precomputed and reused across different queries, reducing query-specific computation and improving scalability.

3.3 Sparse-Evidence Utility and Keyframe Allocation

Sparse-Evidence Utility. Averaging relevance over an entire event can reduce the contribution of a small number of decisive frames. Therefore, we evaluate each event using three relevance statistics. Let $k_e = \max(1, \lceil \rho L_e \rceil)$, where $\rho \in (0, 1]$ and $L_e = a_e - a_{e-1}$ and let $\mathcal{T}_e$ contain the k_e indices with the highest ITM scores in $\mathcal{E}_e$. We compute

$$r_e^{\text{peak}} = \max_{i \in \mathcal{E}_e} r_i, \quad r_e^{\text{top}} = \frac{1}{k_e} \sum_{i \in \mathcal{T}_e} r_i, \quad r_e^{\text{mean}} = \frac{1}{L_e} \sum_{i \in \mathcal{E}_e} r_i. \tag{3}$$

The *Sparse-Evidence Utility* is $u_e = 0.45\, r_e^{\text{peak}} + 0.45\, r_e^{\text{top}} + 0.10\, r_e^{\text{mean}}$. This combination preserves the contribution of frames with high query relevance during keyframe allocation, even when such evidence is sparse within an event.

Keyframe Allocation. To maintain sufficient coverage across events, we introduce a coverage prior that considers both event duration and visual variation: $c_e = \sqrt{L_e}\,(0.75 + 0.25\,\widetilde{v}_e)$, where

$$v_e = \frac{1}{L_e - 1} \sum_{i=a_{e-1}+1}^{a_e - 1} \frac{1 - \mathbf{x}_i^\top \mathbf{x}_{i+1}}{2}, \qquad \widetilde{v}_e = \mathcal{N}\big(\{v_j\}_{j=1}^{M}\big)_e, \tag{4}$$

where v_e measures the average visual change between adjacent frames, and $\widetilde{v}_e$ denotes min–max normalization. We first normalize the event utility and coverage prior across events, and then compute the allocation weight as

$$w_e = \lambda\, \widehat{u}_e + (1 - \lambda)\widehat{c}_e, \tag{5}$$

where $\widehat{u}_e$ and $\widehat{c}_e$ denote the normalized sparse-evidence utility and coverage prior, respectively, and λ controls their relative contributions. The real-valued allocation is $z_e = K w_e$, and the initial event-wise budget is $m_e = \min(L_e, \lfloor z_e \rfloor)$. The remaining keyframes caused by integer rounding are iteratively assigned to events with residual capacity according to the allocation residual $z_e - m_e$, until $\sum_{e=1}^{M} m_e = K$.

3.4 Adaptive Selection

Given an event $\mathcal{E}_e$ with an allocated keyframe budget m_e, we first construct a within-event relevance distribution from the frame-level relevance scores and compute its normalized relevance entropy H_e. The entropy is then used to adapt both the candidate pool size and the relevance–diversity balance according to the concentration of the relevance distribution. Based on the resulting candidate pool, we initialize the highest-relevance frame as an anchor and iteratively select the remaining frames by jointly considering query relevance, visual diversity, and temporal diversity, as summarized in Algorithm 2 of Appendix A.3.

A key component of this stage is the *relevance entropy* H_e, which characterizes the within-event relevance distribution. A lower entropy indicates a more concentrated relevance distribution, whereas a higher entropy indicates that relevance is more dispersed across frames. We use H_e to adapt both the candidate pool size and the relevance–diversity balance. When the entropy is low, a smaller candidate pool is retained and a larger weight is assigned to the relevance term. As the entropy increases, the candidate pool is expanded and more weight is assigned to visual and temporal diversity. Consequently, the candidate pool size h_e varies with H_e rather than remaining fixed. This reduces the risk of discarding relevant frames when relevance is dispersed, while keeping the candidate pool compact when relevance is concentrated.

Proposition 1 (Entropy bounds on candidate retention)**.** *Consider an event with $L_e \geq 2$ frames. Let $\mathbf{p}_e$ denote the within-event softmax distribution used in Algorithm 2, with normalized entropy $H_e = -\sum_{i \in \mathcal{E}_e} p_{e,i} \log p_{e,i} / \log L_e$. For a candidate pool $\mathcal{C}_e$ formed by retaining the h_e highest-relevance frames, define its retained relevance mass as $A_e(h_e) = \sum_{i \in \mathcal{C}_e} p_{e,i}$. For $1 \leq h_e < L_e$,*

$$\max\left\{\frac{h_e}{L_e}, 1 - \frac{H_e \log L_e}{\log(h_e + 1)}\right\} \leq A_e(h_e) \leq \min\left\{1, \frac{h_e}{L_e} + \sqrt{\frac{(1 - H_e)\log L_e}{2}}\right\}. \tag{6}$$

When $h_e = L_e$, $A_e(h_e) = 1$.

For fixed L_e and h_e, lower entropy yields a stronger lower bound on the retained relevance mass. As H_e increases, query relevance becomes more dispersed across frames. In this case, a small candidate pool may retain only a limited fraction of the overall relevance mass, motivating a larger candidate pool. The proof is provided in Appendix A.4.

4 Experiments

4.1 Implementation Details

We sample candidate frames at 1 FPS and use the pretrained BLIP-large image–text matching model (Li et al., 2022) to extract frame-level visual features and ITM scores. For *Semantic Event*

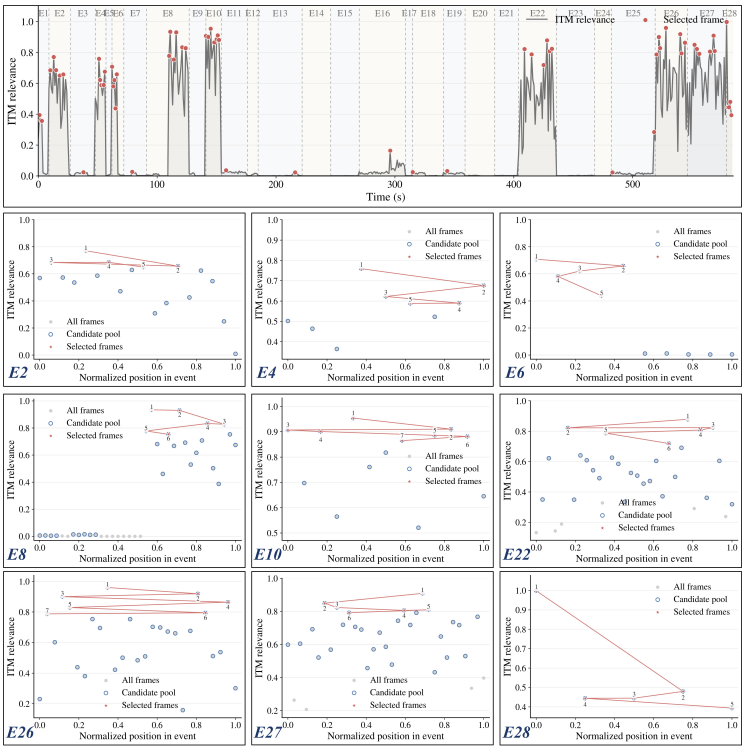

Figure 3: Visualization of *Adaptive Selection*. The upper panel shows the frame-level relevance over the entire video. The lower panels show representative events with different relevance distributions. The visualization illustrates how relevance entropy adapts the candidate pool and the relevance–diversity balance across events.

Partition, we set the minimum and maximum event lengths to $L_{min} = 5$ and $L_{max} = 40$, respectively, the boundary percentile to $P_b = 0.85$, the contextual window size to $w = 3$, the adjacent-frame weight to $\alpha = 0.5$, and the smoothing kernel size to $s = 3$. For *Sparse-Evidence Utility*, the top-relevance ratio is set to $\rho = 0.30$. For event-wise keyframe allocation, we set $\lambda = 0.9$, with the sparse-evidence utility receiving the dominant weight and the coverage prior serving as an auxiliary term. The values of P_b and ρ are determined through hyperparameter sensitivity analysis, as discussed in Appendix A.6 and Appendix A.7, respectively. For *Adaptive Selection*, we set the candidate multiplier to $\kappa = 4$, the entropy temperature to $\tau = 0.07$, and the adaptive relevance weight range to $[\lambda_{\min}, \lambda_{\max}] = [0.55, 0.80]$. The temporal component accounts for $\beta = 0.20$ of the non-relevance weight. The total keyframe budget K follows the setting of each experiment and is varied separately in the frame-budget analysis.

4.2 Comparison with State-of-the-Art Methods

We compare our proposed method with a broad range of frame selection approaches on LongVideoBench, Video-MME, and MLVU using three representative MLLM backbones. For Uniform, Top-k, AKS (Tang et al., 2025), Q-Frame (Zhang et al., 2025a), OneClip-RAG (Chen et al., 2025), BOLT (Liu et al., 2025), FRAG (Huang et al., 2025), E-VRAG (Xu et al., 2025), and QCA (Peng et al., 2026), we directly adopt the results reported in QCA. We further reproduce three recent methods, including WFS-SB (Chen et al., 2026), FOCUS (Zhu et al., 2026), and KFC (Fang et al., 2026), under our evaluation setting. The results are reported in Tab. 1. Our method consistently improves over uniform frame sampling across all three MLLMs and benchmarks. Specifically, the improvements range from 2.6% to 6.7%, with an average gain of 4.14% over all combinations. These results reveal that our method provides more effective visual evidence than uniform temporal sampling.

Table 1: Comparison with state-of-the-art frame selection methods on LongVideoBench, Video-MME, and MLVU. † denotes results reported in QCA, while ‡ denotes results reproduced by us under our evaluation setting. ↑ denotes the improvement over Uniform, and ↑ denotes the improvement over QCA.

Method	LLM Size	LongVideoBench	Video-MME	MLVU
LLaVA-Video-7B				
w/ Uniform†	7B	58.9	64.4	70.8
w/ Top-k†	7B	61.6	63.7	72.8
w/ AKS†	7B	62.7	65.3	71.8
w/ Q-Frame†	7B	61.5	64.7	72.9
w/ OneClip-RAG†	7B	62.5	65.2	71.2
w/ BOLT†	7B	62.2	64.6	70.3
w/ FRAG†	7B	60.6	63.7	69.2
w/ E-VRAG†	7B	63.1	65.4	70.2
w/ QCA†	7B	62.9	66.1	74.1
w/ WFS-SB‡	7B	63.5	66.3	**75.4**
w/ FOCUS‡	7B	63.2	65.7	73.7
w/ KFC‡	7B	62.4	65.6	72.3
w/ **SEEK (Ours)**	7B	**64.2** (↑5.3) (↑1.3)	**67.7** (↑3.3) (↑1.6)	74.9 (↑4.1) (↑0.8)
InternVL3.5-8B				
w/ Uniform†	8B	61.3	61.9	69.9
w/ Top-k†	8B	62.5	61.6	70.4
w/ AKS†	8B	62.9	62.8	70.5
w/ Q-Frame†	8B	62.5	61.7	70.6
w/ OneClip-RAG†	8B	62.1	61.7	70.2
w/ QCA†	8B	63.5	63.9	71.3
w/ WFS-SB‡	8B	63.8	63.5	71.2
w/ FOCUS‡	8B	63.1	63.7	71.4
w/ KFC‡	8B	63.3	63.1	70.7
w/ **SEEK (Ours)**	8B	**64.3** (↑3.0) (↑0.8)	**64.8** (↑2.9) (↑0.9)	**72.5** (↑2.6) (↑1.2)
Qwen3-VL-8B				
w/ Uniform†	8B	63.1	67.6	71.0
w/ Top-k†	8B	64.4	67.4	74.2
w/ AKS†	8B	64.7	68.6	74.2
w/ Q-Frame†	8B	65.8	67.9	74.7
w/ OneClip-RAG†	8B	65.3	68.5	71.3
w/ QCA†	8B	66.9	69.5	75.7
w/ WFS-SB‡	8B	66.0	68.8	75.3
w/ FOCUS‡	8B	67.2	69.2	75.6
w/ KFC‡	8B	64.9	68.7	74.8
w/ **SEEK (Ours)**	8B	**69.8** (↑6.7) (↑2.9)	**71.1** (↑3.5) (↑1.6)	**76.9** (↑5.9) (↑1.2)

Compared with QCA, our method also achieves consistent improvements across all nine settings. For instance, with Qwen3-VL-8B, the gains are 2.9%, 1.6%, and 1.2%on LongVideoBench, Video-MME, and MLVU, respectively. QCA partitions a video into fixed temporal segments, which may split a coherent event or mix semantically different content within the same segment. In contrast, our method constructs variable-length semantic events according to visual transitions, reducing semantic inconsistency and enabling more accurate event-level keyframe allocation.

WFS-SB is particularly relevant to our method because both approaches exploit semantic structure before keyframe selection. However, **SEEK** differs from WFS-SB in both event construction and frame allocation. WFS-SB detects boundaries from query–frame similarity signals and performs Maximal Marginal Relevance (MMR)-based selection within the resulting clips, whereas **SEEK** constructs query-independent variable-length events from local and contextual visual changes, allowing the event partition to be precomputed and reused across different queries. For event-wise allocation, **SEEK** further characterizes sparse evidence using peak relevance, top-k mean relevance, and mean relevance, rather than relying only on clip-level relevance information. Within each event, **SEEK** further uses relevance entropy to adapt the candidate pool and the relevance–diversity bal-

ance to the underlying relevance distribution. These differences enable **SEEK** to better preserve sparse decisive evidence across heterogeneous events.

4.3 Ablation Study

We conduct ablation experiments with InternVL3.5-8B on all three benchmarks to examine the contribution of each component in **SEEK**. As shown in Tab. 2, the full method achieves the best performance, indicating that the proposed components contribute complementary improvements to the final frame selection.

Table 2: Ablation study with InternVL3.5-8B.

Variant	LVB	Video-MME	MLVU	Avg.
w/o Semantic Event Partition	64.2	64.3	71.4	66.6
w/o Sparse-Evidence Utility	63.1	64.4	71.3	66.3
w/o Rel. & Temp. Diversity	63.9	64.6	72.0	66.8
w/o Entropy Adaptation	64.0	64.2	72.2	66.8
SEEK	**64.3**	**64.8**	**72.5**	**67.2**

Replacing *Semantic Event Partition* with fixed temporal segmentation reduces the performance from 64.3% to 64.2% on LongVideoBench, from 64.8% to 64.3% on Video-MME, and from 72.5% to 71.4% on MLVU. The larger decrease on MLVU suggests that the proposed semantic event partitioning is particularly useful when the evidence is distributed across different semantic events. Fixed segmentation may split a coherent event or combine semantically different content, which can make subsequent event-level keyframe allocation less reliable.

Replacing *Sparse-Evidence Utility* with mean relevance causes the largest average performance drop, decreasing the average accuracy from 67.2% to 66.3%. This result supports our motivation that event-level mean relevance is insufficient for capturing sparse decisive evidence.

To examine the contribution of the joint relevance–diversity selection, we replace the full selection score with visual diversity alone by removing the relevance term $\lambda_e^r \widetilde{r}_{e,i}$ and the temporal diversity term $\lambda_e^t D_t(i, \mathcal{S}_e)$. This decreases performance by 0.4%, 0.2%, and 0.5% on LongVideoBench, Video-MME, and MLVU, respectively. This indicates that selecting visually diverse frames alone is insufficient, because visual difference measures appearance variation rather than the semantic evidence required by the query.

Finally, we examine the effect of entropy-based adaptation. For the **w/o Entropy Adaptation** variant, we replace the entropy-dependent candidate pool size and relevance weight with fixed settings: $h_e = \min\{L_e, \max[m_e, \lceil m_e\kappa \rceil]\}$ and $\lambda_e^r = (\lambda_{\min} + \lambda_{\max})/2$. This reduces performance by 0.3%, 0.6%, and 0.3% on LongVideoBench, Video-MME, and MLVU, respectively. The results confirm the benefit of entropy-guided adaptation: a more concentrated relevance distribution yields lower entropy, leading to a smaller candidate pool and a larger relevance weight, whereas a more dispersed distribution yields higher entropy, expanding the candidate pool and placing greater emphasis on visual and temporal diversity.



Figure 4: Visualization of *Semantic Event Partition*. The temporal change signal and representative transitions around detected event boundaries are shown.

4.4 Analysis of Semantic Event Partition

Fig. 4 visualizes the temporal change signal and the resulting semantic event boundaries. The detected boundaries appear at different temporal positions, dividing the video into variable-length events rather than fixed temporal intervals. This allows the partition to better follow the underlying semantic structure of the video. In addition, large peaks in the temporal change signal correspond to noticeable changes in visual content, indicating that the proposed boundary score can effectively

identify semantic event boundaries. In particular, this process requires neither an additional boundary detection network nor query information. It depends only on visual features and can be computed once for each video and reused across different queries, making *Semantic Event Partition* a lightweight and plug-and-play module.

4.5 Analysis of Adaptive Selection

Fig. 3 visualizes *Adaptive Selection* under different relevance distributions. The relevance distribution determines the relevance entropy, which in turn adapts both the candidate pool size and the relevance–diversity balance. When relevance is more concentrated, the lower entropy leads to a smaller candidate pool and a larger relevance weight, directing the selection toward frames with higher relevance. When relevance is more dispersed, the higher entropy expands the candidate pool and assigns more weight to visual and temporal diversity, reducing the risk of discarding relevant frames distributed across the event. The selection patterns further show that the selected frames do not simply follow relevance ranking. Frames with relatively lower relevance can still be selected when they provide additional visual or temporal diversity. This allows *Adaptive Selection* to account for both relevance and diversity rather than relying on frame-level relevance alone.

4.6 Counterfactual Analysis

To examine whether the selected keyframes provide useful evidence for the final answer, we conduct a progressive counterfactual replacement analysis on LongVideoBench using InternVL3.5-8B. Specifically, we progressively replace 25%, 50%, and 75% of the selected keyframes with alternative frames. For each non-zero replacement ratio, we construct three different replacement sets and independently evaluate the video MLLM on each of them. The reported results are averaged over these three runs, while the 0% setting corresponds to the original selected keyframes without replacement. We measure the *ground-truth support margin* of a selected frame set $\mathcal{S}$ as $G(\mathcal{S}) = z_{\mathrm{GT}}(\mathcal{S}) - \max_{c \neq \mathrm{GT}} z_c(\mathcal{S})$, where $z_{\mathrm{GT}}(\mathcal{S})$ denotes the score of the ground-truth answer and $z_c(\mathcal{S})$ denotes the score of an alternative answer. A larger $G(\mathcal{S})$ indicates stronger support for the ground-truth answer. As shown in Fig. 5, progressively replacing the selected keyframes leads to consistent decreases in accuracy, support margin, and ground-truth probability. Restoring the original keyframes also corrects more previously incorrect predictions as the replacement ratio increases. These results show that the frames selected by **SEEK** provide effective evidence for the long-video understanding.

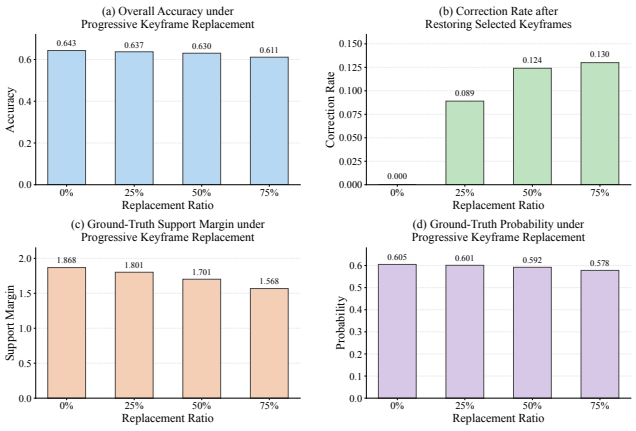

Figure 5: Progressive counterfactual replacement analysis on LongVideoBench with InternVL3.5-8B. Replacing selected keyframes reduces accuracy, support margin, and ground-truth probability, showing that SEEK-selected frames provide effective evidence.

5 Conclusion

We present **SEEK** for long-video keyframe selection, with a focus on preserving sparse evidence across heterogeneous semantic events. **SEEK** combines *Semantic Event Partition* to construct variable-length events, *Sparse-Evidence Utility* to guide event-wise keyframe allocation, and *Adaptive Selection* to balance query relevance with visual and temporal diversity according to relevance entropy. Across three benchmarks and three MLLM backbones, **SEEK** improves over uniform sampling by 2.6%–6.7% and consistently surpasses QCA by 0.8%–2.9% across all nine evaluation settings.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring mandatory disclosure. Generative AI tools were used only for English-language polishing and LaTeX formatting. All AI-assisted edits were reviewed by the authors to ensure technical accuracy and consistency with the original content. We take full responsibility for the final content of this work.

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.

Xinye Cao, Hongcan Guo, Jiawen Qian, Guoshun Nan, Chao Wang, Yuqi Pan, Tianhao Hou, Xiaojuan Wang, and Yutong Gao. VideoMiner: Iteratively grounding key frames of hour-long videos via tree-based group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23773–23783, 2025.

Tao Chen, Shaobo Ju, Qiong Wu, Chenxin Fang, Kun Zhang, Jun Peng, Hui Li, Yiyi Zhou, and Rongrong Ji. Towards effective long video understanding of multimodal large language models via one-shot clip retrieval. *arXiv preprint arXiv:2512.08410*, 2025.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based frame selection by detecting semantic boundary for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24052–24061, 2026.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. Scaling video-language models to 10K frames via hierarchical differential distillation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9964–9981. PMLR, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Bo Fang, YuXin Song, Haoyuan Sun, Qiangqiang Wu, Wenhao Wu, and Antoni B. Chan. Threading keyframe with narratives: MLLMs as strong long video comprehenders. In *International Conference on Learning Representations*, 2026.

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.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-LLM based video frame selection for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13702–13712, 2025.

De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. Frag: Frame selection augmented generation for long video and long document understanding. *arXiv preprint arXiv:2504.17447*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.

Pengyi Li, Irina Abdullaeva, Alexander Gambashidze, Andrey Kuznetsov, and Ivan Oseledets. Maxinfo: A training-free key-frame selection method using maximum volume for enhanced video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7198–7207, 2026.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984. Association for Computational Linguistics, 2024.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3318–3327, 2025.

Jun Peng, Baiyang Song, Jie Li, Hui Li, Yiyi Zhou, Rongrong Ji, and Yonghong Tian. Qca: Query- and content-aware keyframe selection for long video understanding. *arXiv preprint arXiv:2607.00983*, 2026.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16944–16953, 2026.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 54582–54599, 2025.

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.

Shangkun Sun, Ruyang Liu, Haoran Tang, Yixiao Ge, Haibo Lu, Jiankun Yang, and Chen Li. PPLLaVA: Varied video sequence understanding with prompt guidance. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010040.

Wenhui Tan, Xiaoyi Yu, Jiaze Li, Yijing Chen, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. MSJoE: Jointly evolving MLLM and sampler for efficient long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19487–19496, 2026.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

Zeyu Xu, Junkang Zhang, Qiang Wang, and Yi Liu. E-VRAG: Enhancing long video understanding with resource-efficient retrieval augmented generation. *arXiv preprint arXiv:2508.01546*, 2025.

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 (ICCV)*, pp. 22056–22065, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. URL https://openreview.net/forum?id=EElFGvt39K.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 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. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1OQKqLFcbB.

A Appendix

A.1 Additional Related Work

Multimodal Large Language Models for Video Understanding. Video MLLMs extend visual instruction following from images to temporal inputs. Video-LLaVA (Lin et al., 2024) aligns image and video representations before projecting them into the language space, supporting joint multimodal learning. LLaVA-Video (Zhang et al., 2025b) develops a large synthetic instruction dataset covering captioning and question answering to improve video-language learning. More recent backbones further advance multimodal reasoning and long-context processing. InternVL3.5 (Wang et al., 2025) combines cascade reinforcement learning with visual resolution routing, while Qwen3-VL (Bai et al., 2025) incorporates interleaved positional encoding, multi-level visual features, and explicit video timestamps. These developments improve video understanding capabilities, while densely sampled long-video inputs still incur substantial computational cost, motivating efficient visual input reduction.

Efficient Long-Video Understanding. Efficient long-video processing often reduces redundancy at the frame or token level. LongVU (Shen et al., 2025) combines redundant-frame removal, text-guided feature reduction, and spatial compression based on inter-frame dependencies. PPLLaVA (Sun et al., 2026) uses prompt-guided pooling to compress visual sequences while retaining instruction-relevant information. FrameFusion (Fu et al., 2025b) combines similarity-based token merging with importance-based pruning within the model. ViLAMP (Cheng et al., 2025) adopts hierarchical differential distillation, retaining keyframe representations while merging query-salient features from non-keyframes. These methods represent different strategies for reducing visual context. Our work focuses on selecting explicit video frames before downstream inference, without modifying or fine-tuning the video MLLM.

A.2 Evaluation Benchmarks and Evaluation Backbones

Evaluation Benchmarks. We evaluate our method on three widely used long-video understanding benchmarks, including Video-MME (Fu et al., 2025a), MLVU (Zhou et al., 2025), and LongVideoBench (Wu et al., 2024). Video-MME contains videos with diverse durations and domains, covering short-, medium-, and long-form video understanding. MLVU focuses on long videos ranging from several minutes to hours and evaluates different types of temporal reasoning. In particular, we use its multiple-choice evaluation subset in our experiments. LongVideoBench consists of long-form videos with densely annotated multiple-choice questions, requiring video MLLMs to retrieve and reason over temporally distributed evidence. We report accuracy (%) on all three benchmarks.

Evaluation Backbones. To evaluate the generalizability of the proposed frame selection method, we use three representative open-source MLLMs for long-video understanding, including LLaVA-Video-7B (Zhang et al., 2025b), InternVL3.5-8B (Wang et al., 2025), and Qwen3-VL-8B (Bai et al., 2025). These models differ in model architecture, multimodal training strategy, and long-context capability, providing a diverse evaluation setting. For each MLLM, all competing keyframe selection methods use the same checkpoint and frame budget to ensure a fair comparison. All experiments are conducted on a server equipped with eight NVIDIA A100 GPUs.

A.3 Detailed Adaptive Selection Algorithm

Algorithm 2 summarizes the complete within-event adaptive selection procedure.

A.4 Entropy Bounds on Candidate Retention

Proof of Proposition 1. We omit the event subscript for clarity and write $L = L_e$, $h = h_e$, and $H = H_e$. Let $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(L)}$ denote the relevance probabilities sorted in descending order. Since softmax with positive temperature preserves the ordering of ITM scores, the candidate pool retains the first h probabilities, giving

$$A_h = \sum_{j=1}^{h} p_{(j)}. \tag{7}$$

Lower bound. The average of the h largest probabilities is at least the average over the complete distribution. Therefore,

$$\frac{A_h}{h} \geq \frac{1}{L} \sum_{j=1}^{L} p_{(j)} = \frac{1}{L}, \tag{8}$$

which implies $A_h \geq h/L$.

Furthermore, for every $j > h$,

$$(h+1)p_{(j)} \leq \sum_{\ell=1}^{h+1} p_{(\ell)} \leq 1. \tag{9}$$

Hence, $p_{(j)} \leq 1/(h+1)$ for all probabilities outside the candidate pool. Using the nonnegativity of each entropy term, we obtain

$$\begin{aligned} H \log L &= -\sum_{j=1}^{L} p_{(j)} \log p_{(j)} \\ &\geq -\sum_{j=h+1}^{L} p_{(j)} \log p_{(j)} \\ &\geq \sum_{j=h+1}^{L} p_{(j)} \log(h+1) \\ &= (1 - A_h) \log(h+1). \end{aligned} \tag{10}$$

Rearranging and combining the two lower bounds gives

$$A_h \geq \max\left\{ \frac{h}{L},\ 1 - \frac{H \log L}{\log(h+1)} \right\}. \tag{11}$$

Upper bound. Let $\boldsymbol{\pi}$ denote the uniform distribution over the event, with $\pi_i = 1/L$. Its Kullback–Leibler divergence from $\mathbf{p}$ satisfies

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{p} \| \boldsymbol{\pi}) &= \sum_{i=1}^{L} p_i \log \frac{p_i}{1/L} \\ &= \log L + \sum_{i=1}^{L} p_i \log p_i \\ &= (1 - H) \log L. \end{aligned} \tag{12}$$

For the top-h candidate set $\mathcal{C}$, the difference between its probability mass under $\mathbf{p}$ and under $\boldsymbol{\pi}$ is bounded by their total variation distance. Pinsker's inequality then yields

$$\begin{aligned} A_h - \frac{h}{L} &= \sum_{i \in \mathcal{C}} (p_i - \pi_i) \\ &\leq \mathrm{TV}(\mathbf{p}, \boldsymbol{\pi}) \\ &= \frac{1}{2} \sum_{i=1}^{L} |p_i - \pi_i| \\ &\leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbf{p} \| \boldsymbol{\pi})}{2}} \\ &= \sqrt{\frac{(1-H)\log L}{2}}. \end{aligned} \tag{13}$$

Combining this inequality with $A_h \leq 1$ establishes

$$A_h \leq \min \left\{ 1,\ \frac{h}{L} + \sqrt{\frac{(1-H)\log L}{2}} \right\}. \tag{14}$$

When $h = L$, the candidate pool includes all frames and $A_h = 1$. This completes the proof. □

A.5 Qualitative Analysis

Fig. 6 presents two qualitative comparisons between QCA and **SEEK**. In the first example, the question asks for the position of a specific picture on a screen displaying multiple images. Although the two methods select many frames with similar visual content, **SEEK** additionally retains the frames highlighted in green, where the queried picture and its spatial position can be clearly identified. These frames provide direct evidence for locating the picture in the bottom-right region, leading to the correct answer.

The second example shows a similar pattern. The question asks about a later change in the appearance of the woman in the middle. The green-highlighted frames selected only by **SEEK** capture the moment in which the relevant change becomes visible and closely match the query. With this evidence, the MLLM correctly identifies that she tied a black ribbon around her head, whereas QCA produces an incorrect answer.

Compared with QCA, **SEEK** differs in both temporal partitioning and frame allocation. QCA divides the video into fixed temporal segments, which may split a coherent event across segments or group semantically different content within the same segment. In contrast, **SEEK** constructs variable-length semantic events according to visual transitions, allowing keyframe allocation to operate on semantically more coherent temporal units. For frame allocation, QCA distributes frames across segments using query relevance and content variation, whereas **SEEK** jointly considers peak relevance, top-k mean relevance, and average relevance within each event. This makes the allocation more sensitive to sparse but decisive evidence.

A.6 Sensitivity Analysis of Boundary Percentile

We analyze the sensitivity of the boundary percentile P_b, which controls the threshold for identifying semantic event boundaries. Experiments are conducted with InternVL3.5-8B while keeping all other hyperparameters fixed. We evaluate four values, $P_b \in \{0.65, 0.75, 0.85, 0.95\}$, on LongVideoBench, Video-MME, and MLVU.

As shown in Tab. 3, performance generally improves as P_b increases from 0.65 to 0.85. A relatively small P_b produces more event boundaries and may fragment semantically coherent content into short events, which makes subsequent keyframe allocation less reliable. Increasing P_b filters out weaker temporal changes and yields more coherent event partitions, leading to consistent improvements across the three benchmarks.

When P_b is further increased to 0.95, only boundary responses with higher scores are retained, which may overlook semantic transitions with relatively moderate boundary responses and produce



Figure 6: Qualitative comparison between QCA and **SEEK** on two LongVideoBench examples. All displayed frames are selected keyframes. The green boxes highlight frames selected by **SEEK** but not by QCA. These frames contain visual evidence that closely matches the corresponding query and provides critical information for answering the question.

excessively long events. Overall, $P_b = 0.85$ achieves the best balance across the three benchmarks and is therefore used in our experiments.

Table 3: Sensitivity analysis of the boundary percentile P_b on LongVideoBench, Video-MME, and MLVU using InternVL3.5-8B.

P_b	LongVideoBench	Video-MME	MLVU	Avg.
0.65	63.8	64.2	71.7	66.6
0.75	64.2	64.5	72.2	67.0
0.85	**64.3**	**64.8**	72.5	**67.2**
0.95	64.1	64.6	**72.7**	67.1

A.7 Sensitivity Analysis of Top-relevance Ratio

As shown in Tab. 4, **SEEK** exhibits stable performance when ρ varies from 0.25 to 0.35. The differences across the three settings are small on all three benchmarks, indicating that the proposed *Sparse-Evidence Utility* is relatively insensitive to variations in ρ within this range. This is because these settings preserve the contribution of the high-relevance portion of each event while incorporating sufficient frames to obtain a stable estimate of the relevance distribution.

When ρ increases to 0.40, a larger portion of moderately relevant frames is included in the top-relevance term, which weakens its emphasis on sparse decisive evidence and leads to lower performance. Among the tested settings, $\rho = 0.30$ achieves the strongest overall performance and is therefore adopted in our experiments.

Table 4: Sensitivity analysis of the top-relevance ratio ρ on LongVideoBench, Video-MME, and MLVU using InternVL3.5-8B.

ρ	LongVideoBench	Video-MME	MLVU	Avg.
0.25	64.2	64.6	72.2	67.0
0.30	**64.3**	**64.8**	**72.5**	**67.2**
0.35	**64.3**	64.7	72.3	67.1
0.40	64.1	64.3	71.9	66.8

A.8 Effect of Keyframe Budget

Fig. 7 studies the effect of the keyframe budget K using three MLLM backbones on LongVideoBench, Video-MME, and MLVU. We evaluate $K \in \{32, 64, 128\}$ while keeping the remaining settings unchanged.

Across all three MLLMs and three benchmarks, performance consistently improves as the keyframe budget increases. Increasing K allows **SEEK** to retain evidence from a broader range of semantic events and reduces the chance of discarding query-relevant frames under a limited selection budget.

The improvement also varies across MLLMs and benchmarks. InternVL3.5-8B shows relatively smaller gains when increasing the budget from 64 to 128, especially on Video-MME and LongVideoBench, suggesting that much of the relevant evidence has already been retained at $K = 64$. In contrast, LLaVA-Video-7B and Qwen3-VL-8B continue to benefit from the larger budget across all three benchmarks. These results indicate that **SEEK** remains effective under different keyframe budgets, while a larger K provides more complete visual evidence for long-video understanding.

A.9 Additional Visualization of Adaptive Selection

Figure 8 provides additional visualizations of *Adaptive Selection* on two long-video examples.

These examples illustrate the different roles of *Sparse-Evidence Utility* and *Adaptive Selection*. The former operates across events and guides event-wise keyframe allocation, whereas the latter operates

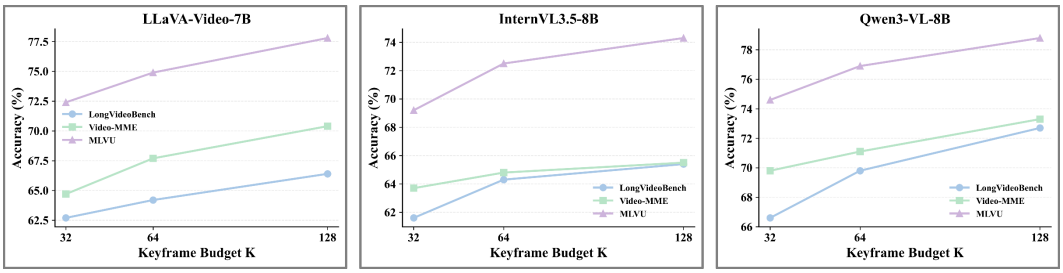

Figure 7: Effect of the keyframe budget K on LongVideoBench, Video-MME, and MLVU with LLaVA-Video-7B, InternVL3.5-8B, and Qwen3-VL-8B. Performance consistently improves as K increases from 32 to 128 across all model–benchmark settings.

within each event and uses relevance entropy to adapt the candidate pool and the relevance–diversity balance. Therefore, an event with low overall relevance does not receive many keyframes simply because its relevance distribution is dispersed, while an event containing stronger sparse evidence can receive a larger allocation and subsequently distribute its selected frames according to the within-event relevance distribution. Overall, the additional visualizations show that SEEK adapts both event-wise allocation and within-event frame selection to different relevance patterns rather than applying a fixed selection strategy across all events.

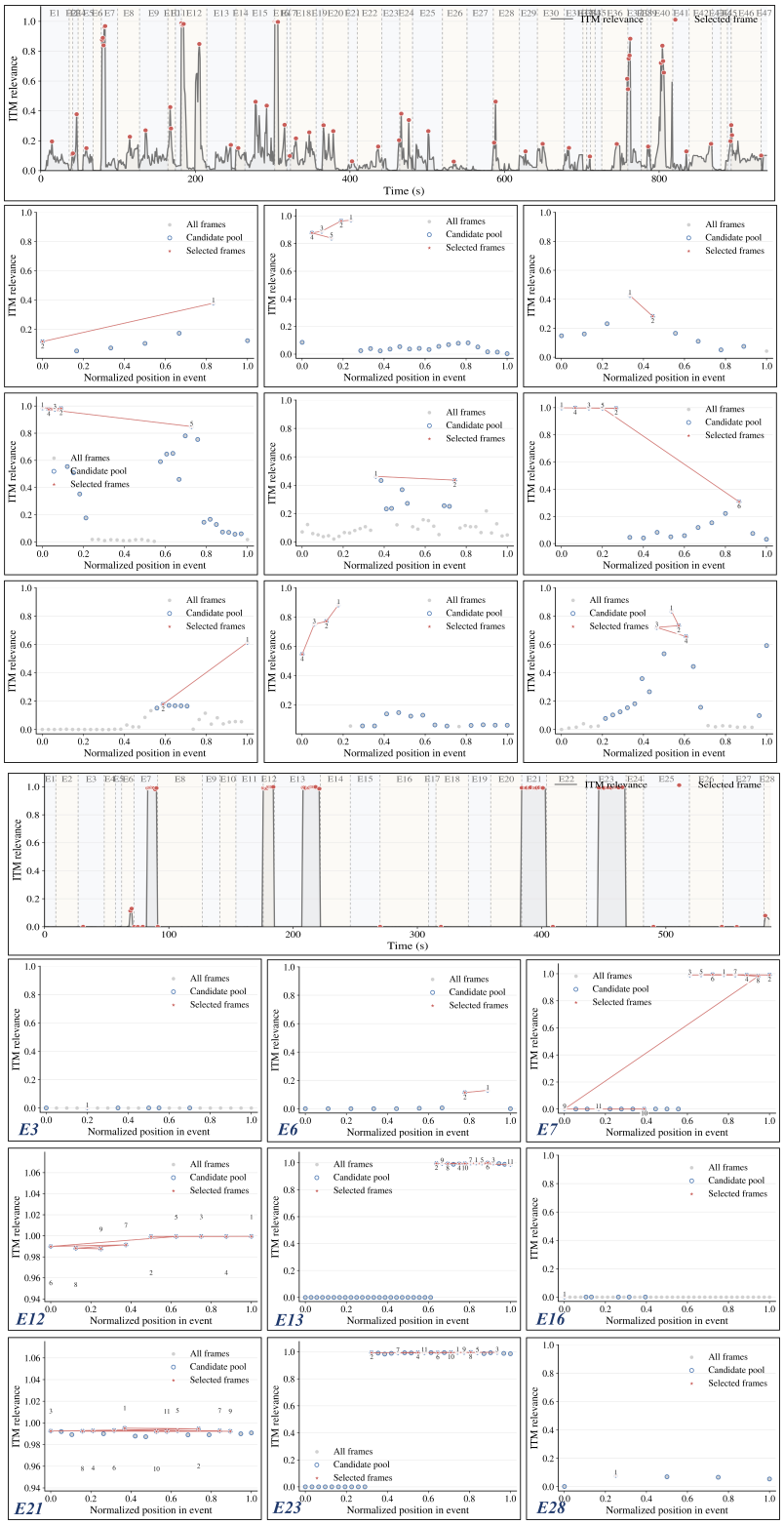

Figure 8: Additional visualization of *Adaptive Selection* on two long-video examples.

Algorithm 2 Adaptive Selection within Event $\mathcal{E}_e$

Require: Event $\mathcal{E}_e$, visual features $\{\mathbf{x}_i\}_{i\in\mathcal{E}_e}$, ITM scores $\{r_i\}_{i\in\mathcal{E}_e}$, keyframe budget m_e
Ensure: Selected keyframes $\mathcal{S}_e$

$L_e \leftarrow |\mathcal{E}_e|$
$r_e^{\text{peak}} \leftarrow \max_{i\in\mathcal{E}_e} r_i$
Compute the relevance distribution

$$p_{e,i} \leftarrow \frac{\exp\big((r_i - r_e^{\text{peak}})/\tau\big)}{\sum_{j\in\mathcal{E}_e}\exp\Big((r_j - r_e^{\text{peak}})/\tau\Big)}$$

Compute relevance entropy

$$H_e \leftarrow -\frac{\sum_{i\in\mathcal{E}_e} p_{e,i}\log p_{e,i}}{\log L_e}, \qquad H_e \leftarrow 0 \text{ if } L_e = 1$$

Compute the adaptive relevance weight

$$\lambda_e^r \leftarrow \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})(1 - H_e)$$

Compute the adaptive candidate pool size

$$h_e \leftarrow \min\{L_e,\, \max[m_e,\, \lceil m_e\kappa(0.85 + 0.45H_e)\rceil]\}$$

$\mathcal{C}_e \leftarrow \text{Top}_{h_e}\{r_i \mid i \in \mathcal{E}_e\}$
if $|\mathcal{C}_e| \le m_e$ **then**
    return $\mathcal{C}_e$
end if
Select the highest-ITM frame as the anchor

$$i^* \leftarrow \arg\max_{i\in\mathcal{E}_e} r_i, \qquad \mathcal{S}_e \leftarrow \{i^*\}$$

Normalize ITM scores within the event and compute

$$\widetilde{r}_{e,i} \leftarrow 0.70\, r_i + 0.30\,\mathcal{N}\left((r_j)_{j\in\mathcal{E}_e}\right)_i$$

Allocate the remaining weight to visual and temporal diversity

$$\lambda_e^v \leftarrow (1 - \lambda_e^r)(1 - \beta), \qquad \lambda_e^t \leftarrow (1 - \lambda_e^r)\beta$$

while $|\mathcal{S}_e| < m_e$ **do**
    for $i \in \mathcal{C}_e \setminus \mathcal{S}_e$ **do**
        Compute visual diversity

$$D_v(i, \mathcal{S}_e) \leftarrow \min_{j\in\mathcal{S}_e} \frac{1 - \mathbf{x}_i^\top \mathbf{x}_j}{2}$$

        Compute temporal diversity

$$D_t(i, \mathcal{S}_e) \leftarrow \frac{\min_{j\in\mathcal{S}_e} |i - j|}{\max(1, L_e - 1)}$$

        Compute the selection score

$$g_e(i \mid \mathcal{S}_e) \leftarrow \lambda_e^r \widetilde{r}_{e,i} + \lambda_e^v D_v(i, \mathcal{S}_e) + \lambda_e^t D_t(i, \mathcal{S}_e)$$

    end for
    Select

$$i^* \leftarrow \arg\max_{i\in\mathcal{C}_e\setminus\mathcal{S}_e} g_e(i \mid \mathcal{S}_e)$$

    $\mathcal{S}_e \leftarrow \mathcal{S}_e \cup \{i^*\}$
end while
return $\mathcal{S}_e$