

When and Where to Look: Adaptive Visual Evidence Scheduling for Efficient Long Video Understanding

Ke Li^{1*}, Jiayu Chen^{2*}, Maoliang Li², Zihao Zheng², Hailong Zou²,
Hengyi Zhang², Xuanzhe Liu², Xiang Chen^{2†}

¹School of Electronics Engineering and Computer Science, Peking University

²School of Computer Science, Peking University

Abstract

Efficient long-video understanding requires vision-language models (VLMs) to reason over a small number of frames selected as sparse visual evidence. Existing relevance-based methods rely on static one-shot selection with fixed frame budgets and candidate pools, while agent-based schedulers achieve adaptivity through costly multi-round reasoning and interactive search. We propose EcoFrame, a training-free framework for low-overhead query-adaptive visual evidence scheduling. EcoFrame leverages the VLM’s inference feedback to determine when to increase the frame budget and where to search for additional candidate evidence. Specifically, entropy-gated budget scheduling uses output uncertainty to stop early when the current evidence is sufficient or progressively expand the frame budget otherwise. Meanwhile, attention-guided candidate proposal converts frame-level attention into a temporal prior, enabling dense local search in informative regions while preserving global coverage when attention is diffuse. Experiments on Video-MME, LongVideoBench, and MLVU demonstrate that EcoFrame achieves a better accuracy-efficiency trade-off across multiple VLM backbones. On Qwen2.5-VL, EcoFrame achieves an average accuracy of 64.4, surpassing BOLT at 63.5, while providing a $1.85\times$ speedup over AKS and BOLT. Compared with the agent-based A.I.R., EcoFrame maintains comparable accuracy with up to a $13.5\times$ inference speedup. Code will be available at <https://github.com/AK-DREAM/EcoFrame>.

1 Introduction

Long-video understanding has become an important capability of modern vision-language models (VLMs), supporting applications such as video question answering and video retrieval. However, directly processing an entire long video remains computationally prohibitive, as a video may contain tens of thousands of frames, far exceeding the context window of VLMs. In practice, only a small subset of frames can be provided as sparse visual evidence. Consequently, selecting informative evidence under a limited frame budget is critical to both inference efficiency and answer accuracy.

A straightforward solution is uniform sampling, which selects frames at fixed temporal intervals. However, this **query-agnostic strategy** overlooks a fundamental property

*These authors contributed equally.

†Corresponding author.

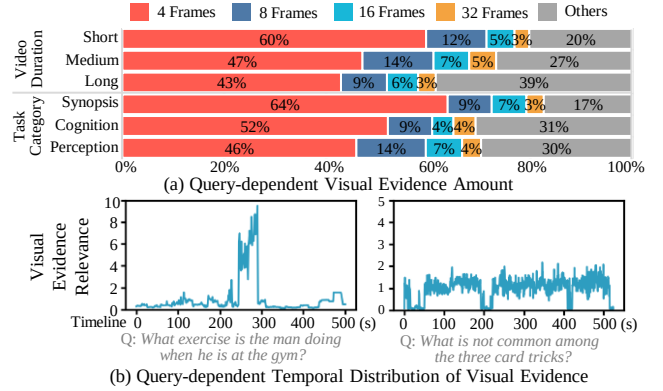


Figure 1: Different queries require different amounts and temporal distributions of visual evidence.

of long-video understanding: different queries require different amounts and temporal distributions of visual evidence, as shown in Fig. 1. A simple query may be answered using only a few frames, whereas a difficult query may require evidence distributed across multiple temporal regions. Therefore, *effective frame selection should adapt both the frame budget and the searched temporal regions to each query*.

Existing **query-aware frame selection methods** (Liu et al. 2025; Zou et al. 2026) can be divided into static and agentic strategies, as illustrated in Fig. 2. *Static methods* typically construct a candidate pool before inference, estimate query-frame relevance using a pretrained vision-language encoder such as CLIP, and select a fixed number of frames in a single step. Although efficient, they cannot adjust the frame budget or candidate pool according to the evidence demand of each query. *Agentic methods*, often implemented through agents, improve adaptivity by iteratively reasoning, verifying candidate frames, and searching for frames. However, their reliance on multiple rounds of VLM reasoning and iterative search introduces considerable scheduling overhead.

The prior methods reveal a fundamental trade-off: static selection is efficient but insufficiently adaptive, whereas agentic selection achieves adaptivity at a high reasoning cost. This raises **the central question of this work**: *how can a VLM adapt its visual evidence to each query without expensive iterative reasoning?* Addressing this problem requires resolving

two coupled challenges. **Challenge 1: Evidence Sufficiency Estimation.** The required amount of evidence varies across queries and is unknown in advance. A fixed frame budget may be insufficient for some queries yet redundant for others. The model must therefore assess whether the current evidence is sufficient and decide *when to stop or expand* the frame budget. **Challenge 2: Targeted Temporal Search.** When the current evidence is insufficient, the model must determine *where to search* for additional frames. A dense candidate pool incurs substantial encoding and relevance-computation costs, whereas a small fixed pool may miss temporally localized evidence. The challenge is to expand the candidate pool toward promising regions without exhaustive processing or costly agent-based search.

To address these challenges with minimal overhead, we analyze two internal signals available during VLM inference: output entropy for answer uncertainty and frame-level attention for temporal focus. This analysis yields two observations. **Observation 1: Output entropy can indicate evidence sufficiency.** Under a limited frame budget, insufficient evidence generally leads to higher output entropy. Entropy can therefore guide *when to exit or expand* the frame budget. **Observation 2: Frame-level attention can indicate where to search.** Frame-level attention correlates with promising temporal regions for candidate expansion: concentrated attention favors denser local search, whereas diffuse attention calls for broader temporal coverage. Attention can therefore guide *where to expand* the candidate pool.

Motivated by these observations, we propose **EcoFrame**, a training-free framework for low-overhead query-adaptive visual evidence scheduling. EcoFrame starts from a sparsely sampled candidate pool, selecting a set of relevant frames for a low-budget answer attempt. Instead of employing a separate agent to make scheduling decisions, it directly converts the VLM’s output entropy and frame-level attention into explicit scheduling signals during inference. **Based on Observation 1**, EcoFrame introduces *entropy-gated budget scheduling*: if the output entropy falls below a round-dependent threshold, the current evidence is considered sufficient and scheduling terminates early; otherwise, the frame budget is progressively expanded. **Based on Observation 2**, EcoFrame introduces *attention-guided candidate proposal*, which propagates frame-level attention into a temporal-cell prior to enable denser search in high-attention regions while maintaining global coverage when attention is diffuse. Input frames are reselected from the expanded candidate pool according to query–frame relevance and temporal coverage, forming coarse-to-fine evidence scheduling.

Extensive experiments on LongVideoBench, VideoMME, and MLVU demonstrate that EcoFrame achieves a favorable accuracy–efficiency trade-off across these benchmarks and generalizes across three distinct VLM backbones, including LLaVA-OneVision, Qwen2.5-VL, and InternVL-3. On Qwen2.5-VL, EcoFrame achieves an average accuracy of 64.4, outperforming the best relevance-based method BOLT at 63.5 while providing a $1.85\times$ speedup over AKS and BOLT. Compared with the agent-based method A.I.R., EcoFrame maintains comparable accuracy while reducing inference latency by up to $13.5\times$.

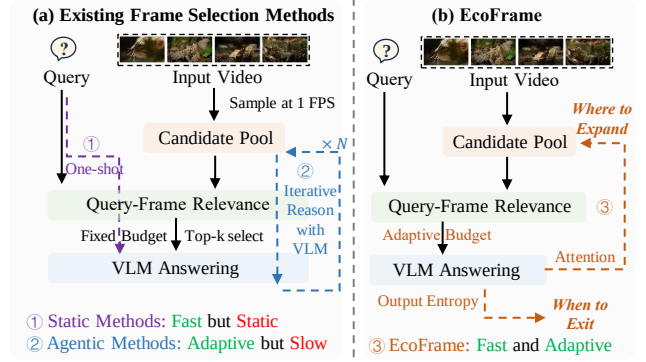


Figure 2: **Frame selection paradigms.** Unlike static methods with fixed budgets and candidate pools or agentic methods with costly iterative reasoning, EcoFrame uses VLM internal signals for efficient visual evidence scheduling.

2 Preliminaries

Relevance-Based Frame Selection. Long-video question answering relies on frame selection to compress a video into a compact set of visual evidence. Formally, given a N -frame video $\mathcal{V} = \{f_i\}_{i=1}^N$ and a query Q , frame selection chooses an input frame set $\mathcal{S} = \{f_{x_i}\}_{i=1}^K \subset \mathcal{V}$, where $K \ll N$, for a downstream VLM to predict $\hat{y} = \text{VLM}(\mathcal{S}, Q)$. The objective is to select sufficient evidence for accurate answering while minimizing the frame budget K .

Uniform sampling selects K frames at fixed intervals and is query-agnostic. Static relevance-based methods instead construct a candidate frame pool $\mathcal{P}$ sampled at a fixed frame rate. They then use a vision–language encoder such as CLIP (Radford et al. 2021) to compute a query–frame relevance score $s_i = \text{sim}(E_v(f_i), E_t(Q))$ for each $f_i \in \mathcal{P}$, and select K frames based on the relevance score. Although query-aware, these methods often fix both $\mathcal{P}$ and K before inference. A dense pool increases encoding and scoring costs, whereas a sparse pool may miss localized evidence; similarly, a fixed frame budget cannot adapt to evidence sufficiency.

VLM Output Entropy. Output entropy characterizes the concentration of the VLM’s predictive distribution. Given an input frame set $\mathcal{S}$ and a query Q , the VLM generates an L -token answer with predictive distribution $p_i(v)$ at position i . Let $\mathcal{V}$ denotes the possible token vocabulary. The token entropy and output entropy are computed as

$$H_i = - \sum_{v \in \mathcal{V}} p_i(v) \log p_i(v), \quad e_{\text{out}} = \frac{1}{L} \sum_{i=1}^L H_i. \quad (1)$$

A lower e_{out} corresponds to a more concentrated distribution and higher predictive certainty, whereas a higher value indicates a more diffuse distribution and greater uncertainty.

Frame-Level Attention Score. During VLM inference, self-attention reflects the model’s focus on visual tokens. Following prior work (Endo, Wang, and Yeung-Levy 2025; Li et al. 2026), we recompute the attention from the last textual query token to all visual tokens before positional embeddings

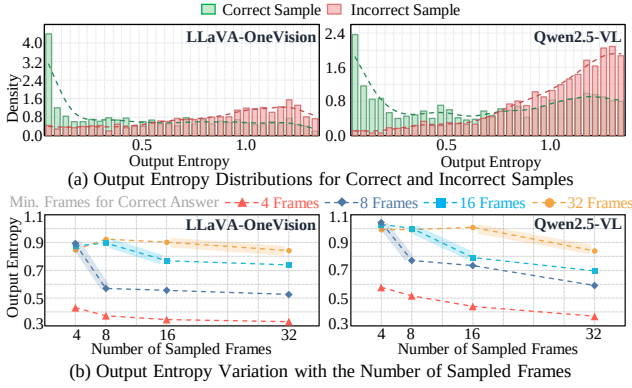


Figure 3: **Output entropy reflects evidence sufficiency.** Correct answers exhibit lower entropy than incorrect ones, with entropy sharply decreasing when the frame budget reaches the minimum required for a correct answer.

are applied. Let $\mathbf{q}_{\ell,h}$ and $\mathbf{K}_{\ell,h}^{\text{Vis}}$ denote the query vector and visual-token key matrix at head h of layer ℓ . Given a set of selected deep layers $\mathcal{L}$ and number of attention heads H , the token-level attention vector $\mathbf{w}$ is computed as

$$\mathbf{w}_{\ell,h} = \text{softmax} \left(\frac{\mathbf{q}_{\ell,h}^\top \mathbf{K}_{\ell,h}^{\text{Vis}}}{\sqrt{D}} \right), \quad \mathbf{w} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{1}{H} \sum_{h=1}^H \mathbf{w}_{\ell,h}, \quad (2)$$

where D is the head dimension. The frame-level attention score a_i is computed by averaging the entries of $\mathbf{w}$ over the visual tokens of frame f_i , representing the attention assigned by the VLM to each input frame during answer generation. This attention is computed only between a few pairs of tokens in selected layers, adding little overhead while remaining compatible with accelerators such as FlashAttention.

3 Motivation and Analysis

EcoFrame is motivated by a simple question: can feedback from low-budget VLM inference reveal *when* additional visual evidence is needed and *where* it should be sought? We study these two decisions through controlled analyses on Video-MME using LLaVA-OneVision and Qwen2.5-VL.

Observation 1: Output entropy can indicate evidence sufficiency. To investigate whether output uncertainty is related to evidence sufficiency, we first compare the output-entropy distributions of correct and incorrect answers under a limited frame budget of 8. As shown in Fig. 3(a), correct answers from both VLMs are concentrated in the low-entropy region, whereas incorrect answers occur more frequently at medium or high entropy. This suggests that when the output entropy is low, indicating low model uncertainty, the model is more likely to have already acquired sufficient visual evidence to answer the question correctly.

We further examine how entropy changes as more frames are introduced. Specifically, for each video-query pair, we evaluate frame budgets $B = 4, 8, 16, 32$ and define B^* as the smallest budget yielding a correct answer. We group queries by B^* and track entropy across budgets. As shown

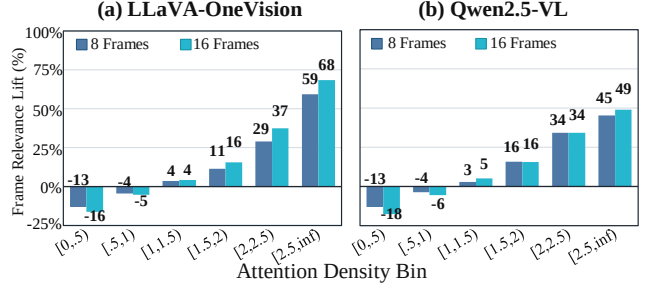


Figure 4: **The correlation between frame-level attention and temporal evidence.** Frame-relevance lift across attention-density bins under 8- and 16-frame inference on (a) LLaVA-OneVision and (b) Qwen2.5-VL. Higher-density bins correspond to more concentrated frame attention.

in Fig. 3(b), for groups with $B^* > 4$, entropy decreases markedly when the budget reaches B^* , i.e., when the model first produces a correct answer after receiving more frames. This pattern is consistent across VLM backbones. Overall, the results reveal that insufficient evidence is generally associated with higher output entropy, while acquiring sufficient evidence coincides with an entropy reduction.

However, absolute entropy levels vary across query groups: even after answering correctly, queries requiring larger B^* retain higher entropy than those resolved with fewer frames. Output entropy therefore indicates evidence sufficiency but is not a universal correctness certificate. It should be interpreted together with the scheduling round to determine *when to stop or expand* the frame budget.

Observation 2: Frame-level attention can indicate where to search. When evidence is insufficient, the problem is where to search for additional frames. To study this, we use the CLIP similarity score between the textual query and all video frames as a proxy for temporal evidence relevance, and compute the frame-level attention score of the input frames to the VLM following Eq. 2. We then group these frames by attention density and compute the relevance lift of their corresponding temporal regions partitioned by adjacent frame midpoints over the video-wide baseline.

Fig. 4 shows a consistent positive relationship between attention concentration and nearby frame relevance under both 8- and 16-frame inference. Temporal regions around diffuse-attention frames with attention density close to 1 exhibit no significant change in relevance. In contrast, the most concentrated attention bin achieves relevance lifts of 59% and 68% on LLaVA-OneVision and 45% and 49% on Qwen2.5-VL under 8- and 16-frame inference, respectively. The rapid increase across attention-density bins indicates that frames with highly concentrated attention are more likely to lie near query-relevant temporal regions.

This association implies a conditional search prior. Concentrated attention suggests that the VLM may have localized a promising region, favoring local search around high-attention frames. Diffuse attention indicates either that no reliable region has yet emerged or that the required evidence is broadly distributed, favoring global coverage.

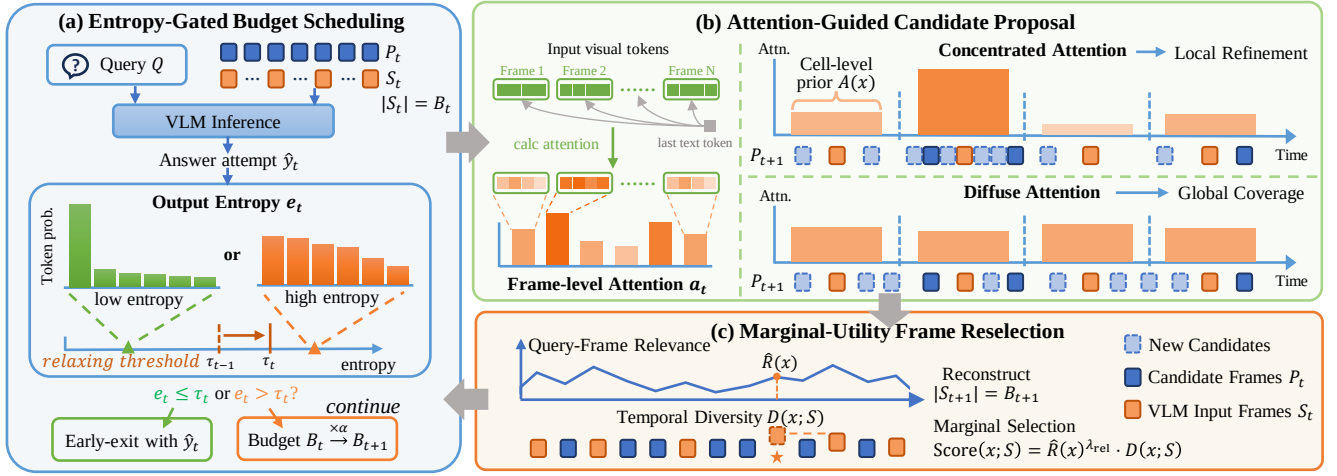


Figure 5: Overview of EcoFrame. Starting from a low-budget input frame set, the target VLM produces an answer, output entropy, and frame-level attention. Entropy controls whether the procedure exits or increases the next-round budget; attention guides candidate pool expansion; relevance and temporal coverage then reconstruct the next input frame set.

4 Method

4.1 Overview

Given a video $V = \{f_i\}_{i=1}^N$ and a query Q , EcoFrame performs query-adaptive visual evidence scheduling over a sequence of frame budgets $B_1 < \dots < B_T \leq B_{\max}$. At round t , it maintains two frame sets with different roles: a candidate frame pool P_t , which stores every frame explored so far together with a cached encoder relevance score, and an input frame set $S_t \subseteq P_t$, which contains exactly B_t frames and is provided to the target VLM. Separating these sets allows the search space to grow incrementally without forcing every explored candidate into the expensive VLM.

Figure 5 summarizes the scheduling method. The VLM first answers the query using S_t , producing answer $\hat{y}_t$, output entropy e_t , and frame-level attention scores $\mathbf{a}_t$. If e_t passes the entropy gate, EcoFrame returns $\hat{y}_t$ immediately. Otherwise, it increases the frame budget to B_{t+1} , uses $\mathbf{a}_t$ as a temporal prior to expand P_t into P_{t+1} , and reconstructs S_{t+1} from the enlarged pool based on query relevance and coverage. This closed loop progressively refines the visual evidence by reusing signals from the target VLM’s own answer attempt, avoiding additional reasoning or verification.

4.2 Entropy-Gated Budget Scheduling

Based on Observation 1, output entropy can serve as a low-overhead evidence-sufficiency signal, guiding *when to stop or expand* the frame budget. Specifically, at the t -th inference round, we compute the output entropy e_t over the generated answer following Eq. 1. A lower e_t indicates a more concentrated predictive distribution, suggesting that the current visual evidence is more likely to be sufficient.

To translate this sufficiency signal into a bounded scheduling procedure, EcoFrame compares e_t with a round-dependent threshold τ_t . If $e_t < \tau_t$, the current evidence is considered sufficient, and scheduling terminates early with

answer $\hat{y}_t$. If $B_t = B_{\max}$, EcoFrame also terminates to ensure a bounded procedure. Otherwise, it progressively expands the frame budget as $B_{t+1} = \min(\lceil \alpha B_t \rceil, B_{\max})$, where $\alpha > 1$. This geometric schedule enables repeated sufficiency estimation at intermediate budgets while reaching $B_{\max}$ in only logarithmically many rounds.

Relaxing Threshold. A global threshold cannot accommodate the query-dependent entropy levels observed in Fig. 3(b): queries requiring more frames often retain higher entropy even after obtaining sufficient evidence. We therefore use a relaxing threshold $\tau_t = \tau_1 + (t-1)\Delta_\tau$ with $\Delta_\tau \geq 0$. Early rounds apply a stricter threshold to prevent premature stopping, whereas later rounds tolerate higher uncertainty of difficult queries, avoiding unnecessary expansion to $B_{\max}$.

4.3 Attention-Guided Candidate Proposal

Based on Observation 2, frame-level attention provides a low-overhead temporal search prior, guiding *where to search* when expanding the candidate pool. Following Eq. 2, we compute the token-level attention vector $\mathbf{w}_t$ during the t -th inference round. To convert $\mathbf{w}_t$ into a frame-level prior that guides candidate proposal, we average $\mathbf{w}_t$ over the tokens of each frame $f_{t,i} \in S_t$ to compute its attention score:

$$a_{t,i} = \frac{1}{|\mathcal{I}_{t,i}|} \sum_{j \in \mathcal{I}_{t,i}} \mathbf{w}_{t,j}, \quad (3)$$

where $\mathcal{I}_{t,i}$ indexes the visual tokens of frame $f_{t,i}$.

Yet, these frame scores exist only on the sparse frame set S_t and cannot directly guide full-video search. We therefore propagate them across the timeline using temporal cells. After sorting the input frames by timestamp, adjacent mid-points define a one-dimensional partition $\{\mathcal{C}_{t,i}\}_{i=1}^{B_t}$. Each unobserved location $x \in \mathcal{C}_{t,i}$ inherits the cell-level prior $A_t(x) = a_{t,i}$, which serves as a timeline-wide search signal.

When the frame budget expands from B_t to B_{t+1} , EcoFrame first determines the number of new candidates as

$M_{t+1} = \lceil m \cdot g(N) \cdot \Delta B_{t+1} \rceil$, where $\Delta B_{t+1} = B_{t+1} - B_t$, m is the candidate expansion ratio, and $g(N)$ is a capped length-aware factor. However, using the propagated prior alone may cluster proposals within high-attention regions. To balance relevance and temporal coverage, an unobserved frame x is scored against the current pool P by

$$D(x; P) = \min_{y \in P} \frac{|x - y|}{N}, \quad (4)$$

$$\text{Score}_{\text{cand}}(x; P) = A_t(x)^{\lambda_{\text{attn}}} \cdot D(x; P), \quad (5)$$

where λ_{attn} controls the sharpness of the attention prior. The attention prior term favors promising regions, while the distance term $D(x; P)$ discourages redundant proposals near existing frames. Since each proposed frame changes the remaining temporal distances, EcoFrame starts from $P := P_t$ and repeatedly adds the highest-scoring frame and updates the distances until M_{t+1} candidates are proposed. The query-frame relevance of these newly proposed candidates are then computed and cached to form P_{t+1} .

Analysis. The multiplicative proposal score yields the desired switch between exploitation and exploration. When attention is concentrated, candidates within high-attention cells receive higher score, inducing dense local search. When attention is diffuse, the cell priors become flatter and the distance term dominates, allocating candidates to poorly covered regions across the video. Intuitively, the number of candidates assigned to each cell approximates

$$\text{Num}_{\text{cand}}(\mathcal{C}_{t,i}) \propto a_{t,i}^{\lambda_{\text{attn}}} \cdot \text{length}(\mathcal{C}_{t,i}), \quad (6)$$

so the same mechanism adapts continuously between local refinement and global coverage.

4.4 Marginal-Utility Frame Reselection

Candidate proposal broadens the inspected region, but the resulting pool generally exceeds the target VLM’s input budget. EcoFrame therefore reconstructs S_{t+1} from the full pool P_{t+1} . We first normalize the cached query-frame relevance $R(x) = \text{sim}(E_v(f_x), E_t(Q))$ of each candidate frame $f_x \in P_{t+1}$ into a $[0, 1]$ distribution: $\hat{R} = \text{softmax}(R/\sigma_R)$, where σ_R is the standard deviation of all $R(x)$.

Selecting by relevance alone may concentrate frames around a single temporal event. To balance relevance and diversity, we measure the marginal utility of adding candidate x to a partially constructed set S as:

$$\text{Score}_{\text{select}}(x; S) = \hat{R}(x)^{\lambda_{\text{rel}}} \cdot D(x; S), \quad (7)$$

where λ_{rel} controls relevance sharpness. Following a similar procedure to candidate pool expansion, EcoFrame initializes S with the most relevant candidate and repeatedly adds the highest-scoring remaining frame until $|S| = B_{t+1}$. Reselecting from the full candidate pool allows newly discovered high-relevance frames to replace weaker earlier frames, while the distance term prevents temporal collapse.

Initialization. Since no frame-level attention prior is available before the first VLM call, EcoFrame uniformly samples an initial pool P_1 of size $M_1 = \lceil m \cdot g(N) \cdot B_1 \rceil$, caches the query-frame relevance scores, and applies the same procedure to construct the initial evidence set S_1 .

Bounded Computation. Although the frame budget grows across rounds, both sources of visual computation remain bounded. The VLM input never exceeds B_{max} , and the geometric schedule bounds cumulative VLM processing by a geometric sum. Since candidate expansion follows the budget increment and $g(N)$ is capped, frame encoding and scoring cost are bounded by $\mathcal{O}(m g_{\text{max}} B_{\text{max}})$. Overall, easy queries terminate with a small budget, whereas difficult queries receive additional but bounded evidence acquisition.

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. We conduct experiments using the lmms-eval (Zhang et al. 2025a) framework on three widely adopted long-video QA benchmarks: Video-MME (Fu et al. 2025), LongVideoBench (Wu et al. 2024), and MLVU (Zhou et al. 2025). For each dataset and VLM backbone, we report average accuracy and end-to-end GPU latency. We also report the average number of frames used for final answering, representing the average frame budget.

Baselines and Models. We compare EcoFrame with a set of representative frame selection methods, including vanilla uniform sampling, static relevance-based selection methods such as AKS, BOLT, and FOCUS (Tang et al. 2025; Liu et al. 2025; Zhu et al. 2026), and an agent-based iterative method A.I.R. (Zou et al. 2026). To assess generalizability across different models, we conduct experiments using three VLMs with diverse architectures: LLaVA-Onevision-7B, Qwen2.5-VL-7B, and InternVL-3-8B (Li et al. 2024; Bai et al. 2025; Zhu et al. 2025).

Implementation Details. For EcoFrame, we initialize the frame budget with $B_1 = 4$ frames and multiply it by $\alpha = 2$ each round. For the early-exit thresholds, we set $\tau_1 = 0.1$ and $\Delta_\tau = 0.2$ for Video-MME and $\tau_1 = 0.2$ and $\Delta_\tau = 0.3$ for LongVideoBench and MLVU to target comparable efficiency regimes. We fix other hyperparameters to $m = 4$, $\lambda_{\text{attn}} = 0.5$, and $\lambda_{\text{rel}} = 1$ across all experiments. To ensure fair comparison, all methods use the same maximum frame budget of 32 and the same CLIP-ViT-L/14 model for query-frame relevance scoring. Relevance-based methods use their default 1 fps candidate pool, and all other hyperparameters follow their official settings. All experiments are conducted on a Linux server with 2 NVIDIA L20 GPUs.

5.2 Main Results

Table 1 reports the main results across different datasets and VLM backbones. Overall, EcoFrame achieves a superior accuracy-efficiency trade-off against the baselines.

Compared with uniform sampling and static relevance-based selection methods, EcoFrame consistently improves average accuracy by up to **4.9%** while using only about 17–21 frames in the final round, and achieves up to **1.89×** speedup over AKS and BOLT. This demonstrates the advantage of query-adaptive evidence scheduling compared to one-shot frame selection. By adapting the frame budget to the evidence demand of each query, EcoFrame avoids redundant

Model	Method	#Frames	Video-MME		LongVideoBench		MLVU		Average		
			Acc. $\uparrow$	Lat. $\downarrow$	Acc. $\uparrow$	Lat. $\downarrow$	Acc. $\uparrow$	Lat. $\downarrow$	Acc. $\uparrow$	Lat. $\downarrow$	Speedup $\uparrow$
LLaVA-OV-7B	Uniform	32	58.6	1.66	56.6	1.67	63.1	1.68	59.4	1.67	–
	AKS	32	59.2	4.50	57.5	3.67	67.6	3.46	61.4	3.88	1.0 $\times$
	BOLT	32	60.4	4.50	58.3	3.67	68.0	3.47	62.2	3.88	1.0 $\times$
	FOCUS	32	58.0	2.92	59.2	2.56	65.0	2.48	60.7	2.65	1.46 $\times$
	A.I.R. [†]	22.3	61.4	28.95	60.7	32.68	<u>69.3</u>	25.98	63.8	29.20	0.13 $\times$
	EcoFrame	19.1	<u>61.0</u>	2.45	<u>59.9</u>	2.19	69.8	1.87	<u>63.5</u>	2.17	1.79$\times$
Qwen2.5-VL-7B	Uniform	32	60.4	1.02	58.9	1.03	59.2	0.99	59.5	1.01	–
	AKS	32	62.4	3.85	57.4	3.04	63.9	2.77	61.2	3.22	1.0 $\times$
	BOLT	32	62.9	3.86	59.3	3.04	<u>68.4</u>	2.76	63.5	3.22	1.0 $\times$
	FOCUS	32	60.0	2.27	58.4	1.90	64.0	1.79	60.8	1.99	1.62 $\times$
	A.I.R. [†]	23.1	65.0	21.03	61.4	25.30	67.5	19.97	64.6	22.10	0.15 $\times$
	EcoFrame	21.3	<u>64.4</u>	1.92	<u>60.3</u>	1.81	68.6	1.49	<u>64.4</u>	1.74	1.85$\times$
InternVL-3-8B	Uniform	32	64.6	1.64	57.9	1.64	67.7	1.75	63.4	1.68	–
	AKS	32	66.2	4.47	58.5	3.65	73.2	3.53	66.0	3.88	1.0 $\times$
	BOLT	32	66.7	4.47	60.4	3.64	73.6	3.53	66.9	3.88	1.0 $\times$
	FOCUS	32	64.2	2.90	59.3	2.55	69.5	2.54	64.3	2.66	1.46 $\times$
	A.I.R. [†]	22.6	68.2	26.74	62.8	31.04	74.5	23.58	68.5	27.12	0.14 $\times$
	EcoFrame	17.4	<u>68.0</u>	2.48	<u>61.1</u>	2.20	<u>74.3</u>	1.48	<u>67.8</u>	2.05	1.89$\times$

Table 1: Main results across different datasets and VLMs. We report accuracy (%) and average end-to-end GPU latency (sec). **Bold** and underlined values indicate the best and second-best results, respectively. [†] denotes an agent-based method.

Method	Acc.	Latency (s)		
		Scoring	Inference	End-to-end
Uniform	56.6	–	1.67	1.67
BOLT	58.3	2.01	1.66	3.67 (1.00 $\times$)
FOCUS	59.2	0.90	1.66	2.56 (1.43 $\times$)
EcoFrame	59.9	0.48	1.71	2.19 (1.68$\times$)
exit $\leq$ 8f	–	0.13	0.38	0.51
exit $\geq$ 16f	–	0.81	2.66	3.47

Table 2: Latency breakdown on LongVideoBench (avg. 12 min) with LLaVA-OV-7B. We decompose the GPU latency into candidate frame scoring and VLM inference. Queries are grouped by their used frame budgets when exit.

computation for simple queries while allocating sufficient budget and targeted evidence search for harder ones.

Compared with the agent-based method A.I.R., EcoFrame maintains comparable accuracy while reducing the average latency by up to **13.5 $\times$** . This efficiency gain comes from its lightweight scheduling mechanism: instead of performing explicit multi-step reasoning or costly VLM-based frame verification, EcoFrame reuses the inference-time entropy and attention signals from the target VLM’s low-budget answer attempt to decide whether to expand the frame budget and where to search next. As a result, EcoFrame preserves the adaptivity of closed-loop evidence acquisition while substantially reducing scheduling overhead.

5.3 Efficiency Analysis

Table 2 decomposes the end-to-end latency into candidate frame scoring and VLM inference. Other scheduling operations with negligible overhead are omitted. Compared with uniform sampling, EcoFrame introduces a small additional cost but yields a substantial accuracy improvement. Compared with static relevance-based methods, EcoFrame provides a more adaptive computation allocation: simple queries can exit early with less candidate frame scoring and VLM computation, while hard queries use additional but bounded refinement rounds. Meanwhile, the attention pattern from coarse inference rounds guide adaptive evidence search, reducing the need for dense frame scoring and improves accuracy. Overall, by adaptively allocating computation according to the evidence demand of each query, EcoFrame achieves higher accuracy with lower average latency.

5.4 Ablation Studies

Module Ablation. Table 3 systematically evaluates the contribution of each proposed module. First, entropy-gated budget scheduling achieves a better accuracy–cost trade-off than fixed-budget inference. Compared with always using the maximum 32 frames budget, it substantially reduces latency with only a small accuracy drop, showing that EcoFrame can adapt the frame budget to the evidence demand of each query. Second, attention-guided candidate proposal is crucial for efficient evidence search. A static fixed-rate pool with same average candidate number suffers from non-adaptability and

Variant	Video-MME		LongVideoBench	
	Acc.	Lat.	Acc.	Lat.
Uniform 32 frames	58.6	1.66	56.6	1.67
EcoFrame	61.0	2.45	59.9	2.19
<i>(A) w/o Entropy-Gated Budget Scheduling</i>				
Fixed budget = 8f	57.6	0.79	57.1	0.78
Fixed budget = 16f	59.2	1.72	58.3	1.70
Fixed budget = 32f	61.5	3.79	60.3	3.75
<i>(B) w/o Attention-Guided Candidate Proposal</i>				
Fixed candidate pool	59.0	2.52	58.3	2.01
w/o attention prior	60.7	2.46	59.2	2.18
<i>(C) w/o Marginal-Utility Frame Reselection</i>				
Relevance-only	60.6	2.45	58.7	2.18
Coverage-only	57.6	2.45	56.8	2.19

Table 3: Ablation of components on Video-MME and LongVideoBench with LLaVA-OV-7B.

τ_1/Δ_τ	#Frames	Acc.	Latency
$\tau_1 = 0.1, \Delta_\tau = 0.0$	25.0	61.2	2.95
$\tau_1 = 0.1, \Delta_\tau = 0.2$	21.0	61.0	2.45
$\tau_1 = 0.1, \Delta_\tau = 0.3$	19.1	60.6	2.21
$\tau_1 = 0.2, \Delta_\tau = 0.0$	22.6	60.7	2.66
$\tau_1 = 0.2, \Delta_\tau = 0.2$	19.3	60.4	2.24
$\tau_1 = 0.2, \Delta_\tau = 0.3$	17.2	60.0	1.97

Table 4: Analysis of different early-exit entropy thresholds on Video-MME with LLaVA-OV-7B. Each row reports the average final frame budget, accuracy, and latency.

often misses crucial evidence. In contrast, EcoFrame progressively expands the candidate pool across rounds, and the accuracy gain over the w/o attention prior variant confirms the critical role of attention-guided proposal. Finally, marginal-utility frame reselection outperforms relevance-only or coverage-only selection by ensuring both representativeness and diversity in the selected frame set.

Entropy Thresholds. Table 4 analyzes the impact of different early-exit entropy thresholds. The results show a consistent trade-off: stricter thresholds allocate more budget to each query, leading to higher accuracy but also higher latency. In addition, compared with a fixed threshold where $\Delta_\tau = 0$, progressively relaxing the threshold across rounds achieves a better trade-off, consistent with our observation in Fig. 3 that harder queries tend to retain higher entropy and benefit from more tolerant later-round thresholds.

Hyperparameter Sensitivity. Figure 6 analyzes the sensitivity of EcoFrame to three key hyperparameters. Increasing the candidate expansion ratio m enlarges the candidate pool and improves accuracy at first, but the gain saturates while latency continues to increase. EcoFrame is less sensitive to the attention and relevance sharpness parameters λ_{attn} and λ_{rel} , showing stable accuracy across a reasonable range without careful tuning. We therefore use moderate default values for all experiments. More results are included in appendix.

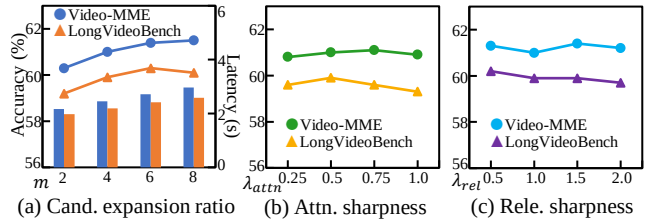


Figure 6: Ablation of candidate expansion ratio m , attention sharpness λ_{attn} , and relevance sharpness λ_{rel} with LLaVA-OV-7B. Accuracy in lines, latency in bars.

6 Related Work

Video Understanding VLMs. Modern vision-language models (VLMs) have evolved from image-centric visual instruction tuning, exemplified by LLaVA (Liu et al. 2023), to video-specialized systems such as Video-LLaVA and Video-ChatGPT (Lin et al. 2024; Maaz et al. 2024), as well as general-purpose multimodal backbones including LLaVA-OneVision, Qwen2.5-VL, and InternVL3 (Li et al. 2024; Bai et al. 2025; Zhu et al. 2025). To scale beyond short clips, recent works explore context and memory compression for VLMs (Song et al. 2024; Zhang et al. 2024; Shen et al. 2025; Shu et al. 2025). However, long videos often contain far more frames than can be densely processed under practical computation and memory budgets. Consequently, extracting a compact set of essential visual evidence remains critical to both inference efficiency and answer accuracy.

Frame Selection for Long-Video Understanding. To fit long videos within VLM context windows, relevance-based frame selection methods improve over query-agnostic uniform sampling by scoring the query relevance of candidate frames using pretrained encoders such as CLIP (Sun et al. 2025; Zhang et al. 2025b; Chen et al. 2026; Ma et al. 2026). AKS and BOLT balance relevance with coverage or diversity (Tang et al. 2025; Liu et al. 2025), while FOCUS and T* reduce scoring costs by exploring promising temporal regions (Zhu et al. 2026; Ye et al. 2025). Nevertheless, these methods often perform one-shot selection from a pre-constructed candidate pool and cannot adapt the frame budget or candidate search based on query-specific evidence requirements and VLM feedback. Agent-based approaches such as VideoAgent and A.I.R. instead use iterative reasoning and evidence verification to guide subsequent searches, improving adaptivity at considerable computation and latency (Wang et al. 2024; Zou et al. 2026; Ding et al. 2026).

7 Conclusion

We introduce EcoFrame, a training-free framework for low-overhead query-adaptive visual evidence scheduling. By using output entropy to adapt the frame budget and frame-level attention to guide candidate expansion, EcoFrame achieves a strong accuracy-efficiency trade-off without expensive agent-based reasoning. Our results demonstrate the value of reusing internal signals produced by the target VLM for lightweight, adaptive evidence scheduling.

References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.
- Chen, W.; Zeng, Y.; Luo, Y.; Xie, T.; Lin, L.; Ji, J.; Zhang, Y.; and Zheng, X. 2026. 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*, 24052–24061.
- Ding, Y.; Lai, X.; Zhang, Y.; Li, W.; Chu, R.; and Yang, Y. 2026. VideoZoomer: Reinforcement-Learned Temporal Focusing for Long Video Reasoning. In *The Fourteenth International Conference on Learning Representations*.
- Endo, M.; Wang, X.; and Yeung-Levy, S. 2025. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22826–22835.
- Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. 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*, 24108–24118.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Li, Y.; Gui, H.; Fan, Z.; Wang, J.; Kang, B.; Chen, B.; and Tian, Z. 2026. Less is more, but where? dynamic token compression via llm-guided keyframe prior. *Advances in Neural Information Processing Systems*, 38: 156861–156904.
- Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, 5971–5984.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, S.; Zhao, C.; Xu, T.; and Ghanem, B. 2025. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3318–3327.
- Lu, Y.; Wang, T.; Rao, F.; Yang, Y.; Zhu, L.; et al. 2026. Flexselect: Flexible token selection for efficient long video understanding. *Advances in Neural Information Processing Systems*, 38: 102751–102777.
- Ma, J.; Zhou, S.; Li, G.; Gao, X.; Cao, Y.; Zeng, H.; Yan, Y.; Wang, Z.; Song, J.; Zheng, B.; et al. 2026. Gift: Global irreplaceability frame targeting for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25610–25620.
- Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 12585–12602.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Shen, X.; Xiong, Y.; Zhao, C.; Wu, L.; Chen, J.; Zhu, C.; Liu, Z.; Xiao, F.; Varadarajan, B.; Bordes, F.; et al. 2025. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. In *International Conference on Machine Learning*, 54582–54599. PMLR.
- Shu, Y.; Liu, Z.; Zhang, P.; Qin, M.; Zhou, J.; Liang, Z.; Huang, T.; and Zhao, B. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26160–26169.
- Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.
- Sun, H.; Lu, S.; Wang, H.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Li, M. 2025. Mdp3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 24090–24101.
- Tang, X.; Qiu, J.; Xie, L.; Tian, Y.; Jiao, J.; and Ye, Q. 2025. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29118–29128.
- Wang, X.; Zhang, Y.; Zohar, O.; and Yeung-Levy, S. 2024. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, 58–76. Springer.
- Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857.
- Ye, J.; Wang, Z.; Sun, H.; Chandrasegaran, K.; Durante, Z.; Eyzaguirre, C.; Bisk, Y.; Niebles, J. C.; Adeli, E.; Fei-Fei, L.; et al. 2025. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8579–8591.
- Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *AAAI*, 9127–9134.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, K.; Li, B.; Zhang, P.; Pu, F.; Cahyono, J. A.; Hu, K.; Liu, S.; Zhang, Y.; Yang, J.; Li, C.; et al. 2025a. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 881–916.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Zhang, S.; Yang, J.; Yin, J.; Luo, Z.; and Luan, J. 2025b. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22056–22065.

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Liang, Z.; Xiao, S.; Qin, M.; Yang, X.; Xiong, Y.; Zhang, B.; et al. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13691–13701.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zhu, Z.; Xu, H.; Luo, Y.; Liu, Y.; Sarkar, K.; Yang, Z.; and You, Y. 2026. FOCUS: Efficient Keyframe Selection for Long Video Understanding. In *The Fourteenth International Conference on Learning Representations*.

Zou, Y.; Jin, S.; Deng, A.; Zhao, Y.; Wang, J.; and Chen, C. 2026. A.I.R.: Enabling Adaptive, Iterative, and Reasoning-based Frame Selection For Video Question Answering. In *The Fourteenth International Conference on Learning Representations*.

A Additional Experimental Results

In this section, we provide additional experimental results that complements the main paper. All experiments use default hyperparameters and LLaVA-Onevision-7B as the VLM backbone if not specified.

A.1 Different Vision–Language Encoders

Table 5 compares different vision–language encoders for frame scoring: CLIP-ViT-B, CLIP-ViT-L (Radford et al. 2021), and SigLIP-so400m (Zhai et al. 2023).

EcoFrame consistently outperforms vanilla uniform sampling across all three encoders, demonstrating its robustness to the encoder choice. Although a smaller encoder like CLIP-ViT-B achieves lower encoding latency, its weaker representations lead to reduced accuracy. We adopt CLIP-ViT-L as the default encoder for its favorable performance.

Encoder	Video-MME		LVB		MLVU	
	Acc.	Lat.	Acc.	Lat.	Acc.	Lat.
Uniform	58.6	1.66	56.6	1.67	63.1	1.68
CLIP-ViT-B	60.4	1.93	58.9	1.79	66.8	1.51
CLIP-ViT-L	61	2.45	59.9	2.19	69.8	1.87
SigLIP-so400m	61.1	3.79	58.7	3.29	69.3	2.78

Table 5: Ablation of different vision–language encoders for computing query–frame relevance scores.

A.2 Different Frame Budget Schedules

Table 6 compares alternative frame budget schedules against our default schedule $B = [4, 8, 16, 32]$.

Reducing the maximum budget from 32 to 16 reduces latency but results in a clear accuracy drop. Starting from 8 frames prevents early exits at smaller budgets and achieves higher accuracy on Video-MME at the cost of increased latency. Directly jumping from 4 to 32 frames reduces the number of refinement rounds, but removes intermediate opportunities to reassess evidence sufficiency and redirect the evidence search, leading to lower accuracy.

Schedule	Video-MME		LongVideoBench	
	Acc.	Lat.	Acc.	Lat.
$B = [4, 8, 16]$	58.7	1.37	58.4	1.31
$B = [8, 16, 32]$	61.4	2.70	59.8	2.55
$B = [4, 32]$	60.8	2.26	59.4	2.13
$B = [4, 8, 16, 32]$	61.0	2.45	59.9	2.19

Table 6: Ablation of different frame budget schedules. The schedule list B denotes the frame budget of each round.

A.3 Length-aware Candidate Number

During candidate pool expansion, the candidate number is partially determined by a capped length-aware factor $g(N)$ that grows sublinearly with the video length N :

$$g(N) = \min(\max(\sqrt{N/N_0}, 1), g_{max}), \quad (8)$$

where we set N_0 to 2 min and $g_{max} = 4$. This allows the candidate number to scale mildly to video length, while avoiding the unbounded scoring cost caused by linear growth.

To validate this design, we compare different length-aware factors in Table 7. A constant factor cannot scale with video length and may miss relevant evidence in longer videos, while linear growth produces denser candidate pools at the cost of substantial redundant scoring. In contrast, capped sublinear growth is sufficient when combined with attention-guided proposal, yielding a better accuracy–efficiency trade-off.

$g(N)$	Video-MME		LongVideoBench	
	Acc.	Lat.	Acc.	Lat.
Constant	60.3	2.14	59.5	1.95
Linear	60.9	4.07	60.1	3.08
Sublinear (Ours)	61.0	2.45	59.9	2.19

Table 7: Comparison of different length-aware factors for determining the candidate number. Constant ($g(N) = 1$), Linear ($g(N) = N/N_0$), Sublinear (Eq. 8).

A.4 Evaluation on Open-Ended Video QA

To assess whether EcoFrame generalizes to open-ended video QA, we further evaluate it on ActivityNet-QA (Yu et al. 2019) using a uniformly sampled subset of 1600 samples from the test split. ActivityNet-QA contains free-form questions over diverse human activities and evaluates the semantic correctness of generated answers. Since the dataset mainly contains shorter videos, we set the maximum frame budget to 16 frames for all methods. We use the evaluation framework integrated in lmms-eval (Zhang et al. 2025a), using gpt-4o-mini as the automatic judge for all VLM responses following official protocol.

As shown in Table 8, EcoFrame achieves a better overall accuracy–efficiency trade-off on ActivityNet-QA, outperforming uniform sampling and BOLT while requiring fewer final input frames on average. Notably, compared with the w/o early exit variant that always proceeds to the maximum budget, the full method achieves comparable accuracy with fewer frames and lower inference latency. This result demonstrates that entropy-gated budget scheduling remains effective in open-ended video QA.

Method	#Frames ↓	Accuracy ↑	Latency ↓
Uniform	16.0	54.5	0.91
BOLT	16.0	55.2	2.2
EcoFrame-noE	16.0	55.5	1.71
EcoFrame	12.9	55.4	1.37

Table 8: Results on the open-ended ActivityNet-QA benchmark with LLaVA-OneVision-7B. All methods use a maximum budget of 16 frames. EcoFrame-noE denotes the w/o early exit variant.

A.5 Additional Analysis on Scheduling Signals

To assess the reliability of output entropy and frame-level attention on other benchmarks, we repeat the same analysis experiments of these signals on LongVideoBench and ActivityNet-QA.

As shown in Fig. 7 and Fig. 8, both benchmarks exhibit patterns consistent with our main observations: lower output entropy is associated with more sufficient visual evidence, while concentrated frame-level attention more reliably identifies temporally relevant regions.

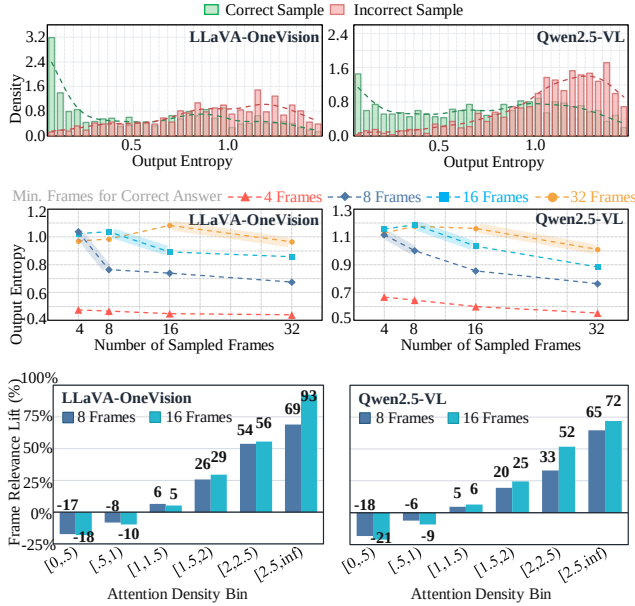


Figure 7: Analysis of output entropy and frame-level attention on LongVideoBench.

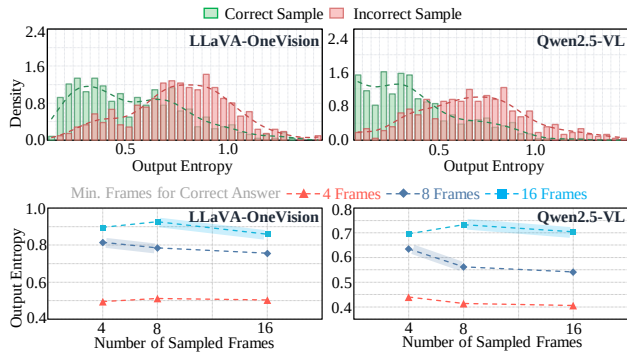


Figure 8: Analysis of output entropy on ActivityNet-QA.

B Frame-Level Attention Extraction

Previous studies have explored using VLM attention to estimate the importance of visual tokens (Chen et al. 2024; Li et al. 2026). FlexSelect (Lu et al. 2026) finds that attention from intermediate-to-deep layers provides the most reliable importance estimates, while Feather (Endo, Wang, and

Yeung-Levy 2025) shows that rotary positional embeddings can introduce positional bias into attention distributions.

Inspired by these findings, we further investigate frame-level attention as a temporal prior for long-video evidence acquisition. Specifically, we compute attention from the query-key representations before applying positional embeddings over a set of selected deep layers, as defined in Eq. 2 of the main paper, and average the resulting token-level scores within each frame. Since all evaluated models have 28 attention layers, we use layers 19–21 as reference layers by default. We ablate different attention extraction schemes and choices of reference layers below.

Attn. Extraction	LLaVA-OV		Qwen2.5-VL	
	V-MME	LVB	V-MME	LVB
w/o attention	60.7	59.2	63.8	59.3
post-RoPE (19–21)	61.0	59.7	64.2	60.1
Layers 0–3	60.9	59.8	64.2	59.4
Layers 4–7	61.0	59.6	64.4	59.9
Layers 8–11	60.8	59.9	64.5	59.5
Layers 12–15	60.9	59.2	64.0	59.6
Layers 16–18	60.9	59.5	64.2	59.8
Layers 19–21	61.0	59.9	64.4	60.3
Layers 22–24	61.0	59.8	64.1	60.1
Layers 25–27	61.2	59.8	64.2	59.8

Table 9: Ablation of frame-level attention extraction. We evaluate attention computed post-RoPE and pre-RoPE attention extracted from different groups of reference layers. We report accuracy on Video-MME and LongVideoBench using LLaVA-Onevision-7B and Qwen2.5-VL-7B.

As shown in Table 9, pre-RoPE attention extracted from certain deep layers such as 19–21 provides an overall more reliable temporal prior, achieving more consistent performance across datasets and models while consistently improving over the w/o attention variant.

C Limitations and Discussion

Scenario-Dependent Compute Configuration. Although EcoFrame enables adaptive evidence scheduling for different queries, the entropy thresholds and maximum frame budget remain globally configured. The most preferable operating point may vary across models, video domains, and deployment requirements. Future work could adapt these compute configurations automatically through lightweight calibration or cost-aware scheduling policies.

Needle-in-a-Haystack Evidence. Extremely short evidence segments may remain challenging when they occupy only a very small fraction of a long video. Attention-guided candidate proposal mitigates this issue by preserving global coverage when the attention distribution is diffuse, but extremely brief events may still be missed with limited candidate frames. Future work could incorporate complementary temporal cues, such as event boundaries or motion saliency, to enable more precise evidence search.