

VISA: Video Skeleton-Aware Efficient Frame Selection for Long Video Understanding

Qihui Zhu^{1*}, Zijian Wen^{1*}, Tao Zhang^{1*}, Mengjie Zhang¹, Shenghao Ye¹,
Dong Jin², Shuangwu Chen^{1†}, Xiaobin Tan¹, Quan Zheng¹, Jian Yang¹

¹University of Science and Technology of China

²Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

{qh.zhu, wzj20020304, zhangtaolqy, zhangmengjie, ssh0321y, kingdon}@mail.ustc.edu.cn

{chensw, xbtan, qzheng, jianyang}@ustc.edu.cn

Abstract

Long-video understanding remains challenging for multimodal large language models (MLLMs), as informative visual evidence is often sparsely distributed across many redundant frames. Although sampling more frames can improve coverage, it also increases visual encoding overhead and consumes more visual tokens, making compact frame selection essential. However, existing methods often sample candidate frames uniformly before selection, which may require decoding most frames of the video. Moreover, simply selecting frames by query relevance may overlook the global temporal structure needed for coherent video understanding. To address these limitations, we propose **VISA**, a **V**ideo **S**keleton-Aware framework for efficient long-video frame selection. VISA exploits sparse I-frame anchors from the native GOP structure to avoid dense candidate decoding and constructs a video skeleton that preserves the coarse temporal progression. Query relevance is aggregated over this skeleton to select informative anchors, followed by cross-GOP refinement to obtain final frames. Extensive experiments show that VISA consistently improves long-video understanding, achieving up to 8.8% relative gain over uniform sampling. Meanwhile, VISA reduces average end-to-end latency by 32.9%–53.6% compared with recent frame selection baselines on LongVideoBench, demonstrating both effectiveness and efficiency for long-video frame selection.

1 Introduction

Multimodal large language models (MLLMs) (Liu et al., 2023; Li et al., 2023; Bai et al., 2023; Lai et al., 2024; Wang et al., 2024a) have recently demonstrated remarkable capabilities in video understanding (Lin et al., 2024; Maaz et al., 2024; Li et al., 2025). However, as video duration increases,

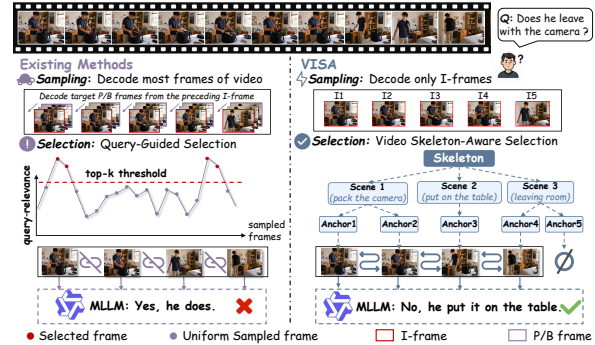


Figure 1: **Comparison of frame selection methods.** Existing methods (Left) decode most video frames and select query-relevant frames, e.g., “*packing the camera*” and “*leaving room*”, while missing the structural evidence “*put on the table*”. Our method (Right) decodes sparse I-frames as anchors to build a video skeleton, and decodes only selected neighboring GOPs for refinement. This design preserves query-relevant evidence and structural completeness with lower decoding cost.

exhaustively processing every frame becomes impractical. A typical long video may contain tens of thousands of frames, each encoded into hundreds of visual tokens. As a result, the accumulated visual tokens substantially increase inference latency and may even exceed the limited context capacity of current MLLMs (Wang et al., 2024b; Li et al., 2024b; Jiang et al., 2025; Shu et al., 2025). To address this challenge, frame selection has emerged as an effective method for reducing video understanding cost by retaining essential visual information with a compact set of key frames.

Existing frame selection methods (Tang et al., 2025; Zhu et al., 2025b; Zhang et al., 2025c; Yao et al., 2025; Chen et al., 2026) commonly follow a two-stage pipeline before MLLM inference: candidate frame sampling followed by key frame selection under a limited budget. The sampling stage often relies on uniform sampling (Liu et al., 2023). However, sampled frames are often P- or B-frames that cannot be decoded independently. As illus-

*Equal contribution.

†Corresponding author.

trated in Fig. 1, recovering them requires decoding all frames sequentially from the preceding I-frame within the same Group of Pictures (GOP), which incurs substantial decoding overhead. In the selection stage, existing methods typically select frames according to query relevance (Tang et al., 2025; Zhang et al., 2025c). However, retaining only query-relevant frames leads to the loss of contextual information, which may break video semantic coherence (e.g., temporal logic and action continuity). As shown in Fig. 1, the missing of implicit clues tends to yield biased reasoning results. Therefore, an effective frame selection strategy should retain not only local query-relevant frames but also global contextual frames, while avoiding unnecessary decoding overhead.

Given the native structure of videos, I-frames capture key object, background, and scene features, can be decoded independently, and serve as references for subsequent frames within a GOP. These properties make I-frames suitable anchors for constructing a video skeleton that captures the global temporal structure and contextual information of the video. Leveraging this video skeleton for key frame selection allows us to decode only sparse I-frames while preserving the video semantics.

Driven by this insight, we propose **VISA**, a **Video Skeleton-Aware** framework for efficient long-video frame selection. VISA directly extracts sparse I-frame anchors from the video’s native GOP structure and groups them into query-agnostic scenes, forming a compact video skeleton that captures the global temporal structure of the video. Query relevance is then aggregated over this hierarchy to capture the critical evidence. Under a constrained budget, VISA selects informative and non-redundant anchors and refines their neighboring GOPs to identify the final frames. In this way, VISA bridges frame-level evidence and skeleton-level cues, capturing both query-relevant and structurally important frames for effective MLLM-based long-video understanding.

Our Contributions. (1) *New Insight.* We leverage I-frames as video-native, decoding-efficient anchors for video skeleton construction, providing structural support for frame selection. (2) *Novel Approach.* We propose **VISA**, a video skeleton-aware frame selection method that builds a video skeleton from I-frame anchors, selects informative anchors, and refines them into final frames. (3) *Strong Performance.* Extensive experiments on multiple MLLMs and benchmarks show that

VISA achieves competitive or better answer accuracy than recent frame selection baselines, while substantially reducing end-to-end latency.

2 Method

Given a long video V , a textual query q , and a frame budget B , VISA aims to select a compact frame set X^* with $|X^*| \leq B$ for MLLM inference, while preserving query-relevant evidence and temporal context with low decoding cost. As shown in Fig. 2, VISA follows a coarse-to-fine frame selection process: it first derives a query-agnostic temporal skeleton from sparse I-frame anchors and enriches it with query relevance through hierarchical evidence aggregation (Sec. 2.1); then allocates the frame budget over the relevance-enriched skeleton to select informative and non-redundant anchors $\mathcal{A}^*$ (Sec. 2.2); and finally refines each selected anchor across neighboring GOPs to obtain the final selected frame set X^* (Sec. 2.3).

2.1 Video Skeleton Construction

I-frame Anchors and Scene Segments. We first construct a query-agnostic video skeleton by leveraging the native GOP structure of the video. Specifically, we extract I-frames as sparse temporal anchors and define the anchor set as $\mathcal{A} = \{a_1, a_2, \dots, a_M\}$, where each anchor a_i corresponds to the I-frame of a GOP together with its temporal position. We further group anchors into coherent scene segments by estimating scene boundaries from the I-frame anchor sequence. For each anchor frame, we convert RGB to a luma map by $Y_i = 0.299R_i + 0.587G_i + 0.114B_i$, resize it to 32×32 , and vectorize and ℓ_2 -normalize it as g_i . The visual change between adjacent anchors is measured by cosine distance $\delta_i = 1 - g_i^\top g_{i+1}$. To reduce spurious local fluctuations, we smooth the change scores with a three-point moving average and detect local peaks using an adaptive threshold, yielding query-agnostic scene segments $\mathcal{S} = \{S_1, S_2, \dots, S_K\}$. These scene segments, together with the I-frame anchors, define a query-agnostic video skeleton that retains the native temporal structure of the video.

Query-aware Evidence Aggregation. After constructing the video skeleton, we inject query information by computing the relevance between the query q and each I-frame anchor a_i . We encode the query and anchor frames using CLIP (Radford et al., 2021), obtaining normalized embeddings e_q

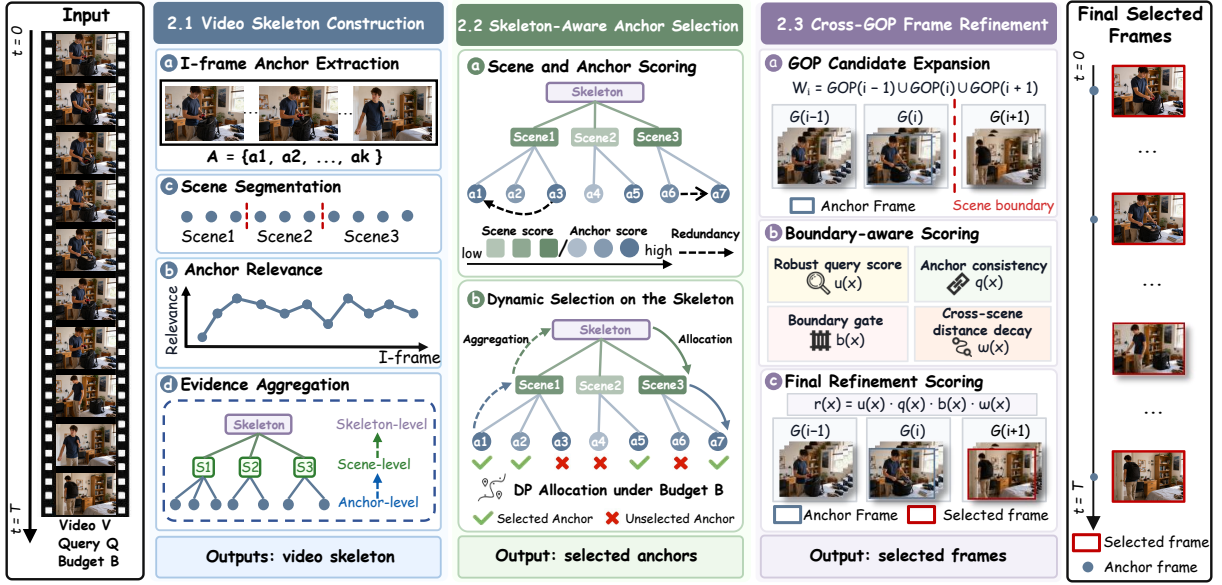


Figure 2: **Workflow of VISA.** VISA performs efficient long-video frame selection through three stages: (1) *Video Skeleton Construction* builds a video skeleton from sparse I-frame anchors; (2) *Skeleton-Aware Anchor Selection* selects informative and non-redundant anchors; and (3) *Cross-GOP Frame Refinement* refines selected anchors within neighboring GOPs to obtain the final selected frames.

and e_i , respectively. The raw query-anchor similarity is computed as $s_i = e_i^\top e_q$. We then apply within-video min-max normalization to calibrate the anchor-level relevance scores. The normalized score p_i denotes the calibrated query relevance of anchor a_i . Given the anchor-level relevance scores, we aggregate query-related evidence from anchors to scenes and then to the whole video skeleton:

$$E_k = \tau \log \left(\frac{1}{|S_k|} \sum_{i \in S_k} \exp(p_i / \tau) \right), \quad (1)$$

$$E_v = \tau \log \left(\frac{1}{K} \sum_{k=1}^K \exp(E_k / \tau) \right).$$

Here, τ is a temperature parameter empirically set to 0.1. This process yields a relevance-enriched video skeleton with anchor-level relevance p_i , scene-level evidence E_k , and video skeleton-level evidence E_v .

2.2 Skeleton-Aware Anchor Selection

Given the relevance-enriched video skeleton from Sec. 2.1, we select a compact set of informative anchors under a fixed budget B . Instead of treating anchor selection as a flat top- B ranking problem over all anchors, we formulate it as a skeleton-aware allocation problem. This formulation explicitly considers both scene-level evidence and anchor-level distinctiveness, encouraging the selected anchors

to cover query-relevant scenes while avoiding redundant anchors within each scene.

Skeleton-aware Objective. Let $\mathcal{A}^* \subseteq \mathcal{A}$ denote the selected anchor set with $|\mathcal{A}^*| = B$. For each scene S_k , we define the selected anchors inside it as $\mathcal{A}_k^* = \mathcal{A}^* \cap S_k$ and $m_k = |\mathcal{A}_k^*|$. We optimize the following skeleton-aware objective:

$$\max_{\mathcal{A}^* \subseteq \mathcal{A}, |\mathcal{A}^*| = B} \sum_{k=1}^K \left[\Delta_k^{\text{scene}} g(m_k) + \sum_{a_i \in \mathcal{A}_k^*} w_i \tilde{\Delta}_i \right], \quad (2)$$

where Δ_k^{scene} measures the evidence surplus of scene S_k , and $\tilde{\Delta}_i$ denotes the effective gain of anchor a_i after redundancy discounting. The function $g(m) = \log(1+m)$ introduces diminishing returns, preventing the budget from being overly concentrated in a single scene. For each selected anchor $a_i \in \mathcal{A}_k^*$, we define a rank-dependent marginal weight $w_i = g(r_i) - g(r_i - 1)$, where r_i is the rank of a_i after sorting anchors in $\mathcal{A}_k^*$ by $\tilde{\Delta}_i$ in descending order. The first term in Eq. 2 encourages allocation to query-relevant scenes, while the second term favors distinctive anchor leaves with diminishing marginal weights.

Scene and Anchor Scoring. We define the scene surplus relative to the skeleton-level evidence and

the anchor surplus relative to its parent scene:

$$\begin{aligned}\Delta_k^{\text{scene}} &= [E_k - E_v]_+, \\ \Delta_i &= [p_i - E_k]_+, \quad a_i \in S_k,\end{aligned}\quad (3)$$

where $[x]_+ = \max(x, 0)$. Here, Δ_k^{scene} indicates whether scene S_k contains stronger query-related evidence than the skeleton-level baseline, while Δ_i measures whether anchor a_i is locally more distinctive than its parent scene.

High-surplus anchors may still be redundant if their temporal neighbors contain similar visual content. We therefore compute a local redundancy score within a temporal radius r :

$$\rho_i^{(r)} = \frac{1}{|\mathcal{N}_r(i)|} \sum_{j \in \mathcal{N}_r(i)} \frac{e_i^\top e_j + 1}{2}, \quad (4)$$

where $\mathcal{N}_r(i)$ denotes the valid temporal neighbors of a_i within radius r , excluding a_i itself. The effective anchor gain is then defined as:

$$\tilde{\Delta}_i = \Delta_i \left(1 - \rho_i^{(r)}\right). \quad (5)$$

This favors anchors that are both query-relevant and visually distinctive. We set $r = 2$ in our final model.

Dynamic Selection on the Skeleton. To solve Eq. 2 efficiently, we decompose it into scene-wise utilities. For each scene S_k , we sort all its anchors by $\tilde{\Delta}_i$ in descending order, and denote the j -th largest value as $\tilde{\Delta}_{k,(j)}$. If m anchors are assigned to scene S_k , the corresponding scene utility is:

$$U_k(m) = \Delta_k^{\text{scene}} g(m) + \sum_{j=1}^m w_j \tilde{\Delta}_{k,(j)}, \quad (6)$$

This is the scene-wise form of Eq. 2, where the selected anchor leaves in each scene are instantiated by the top- m anchors according to $\tilde{\Delta}_i$. The first term preserves query-relevant scene, while the second term rewards high-gain anchor leaves with diminishing marginal weights.

We allocate the global budget using dynamic programming:

$$\text{dp}[k, b] = \max_{0 \leq t \leq \min(b, |S_k|)} [\text{dp}[k-1, b-t] + U_k(t)]. \quad (7)$$

Here, $\text{dp}[k, b]$ denotes the maximum utility obtained from the first k scenes under budget b , and t is the number of anchors assigned to scene S_k . By

backtracking the DP table, we determine the number of anchors assigned to each scene and select the corresponding top-ranked anchors, yielding the final anchor set $\mathcal{A}^*$. The selected anchors are then used as coarse temporal evidence locations for the following neighboring-GOP refinement stage.

2.3 Cross-GOP Frame Refinement

The selected anchors $\mathcal{A}^*$ provide coarse temporal locations of potential informative evidence, but the most informative frames may lie near GOP boundaries or in neighboring GOPs rather than exactly at the anchors. Therefore, for each selected anchor $a_i \in \mathcal{A}^*$, we expand its refinement window to adjacent GOPs:

$$W_i = \text{GOP}(i-1) \cup \text{GOP}(i) \cup \text{GOP}(i+1). \quad (8)$$

While this expansion improves recall for boundary-near evidence, it may also introduce candidates from neighboring scenes. We therefore apply boundary-aware scoring to each candidate frame $x \in W_i$.

Boundary-aware Scoring. For each candidate frame $x \in W_i$, we compute its raw query similarity as $s(x) = e_x^\top e_q$, where e_x is the CLIP image embedding of x . Since score distributions may vary across different local GOP windows, we calibrate the query score using robust statistics:

$$u_i(x) = \frac{1}{1 + \exp\left(-\frac{s(x) - \mu_i}{\sigma_i}\right)}, \quad (9)$$

where $\mu_i = \text{median}_{x \in W_i} s(x)$ and $\sigma_i = 1.4826 \cdot \text{median}_{x \in W_i} |s(x) - \mu_i| + \varepsilon$. The calibrated score $u_i(x)$ measures how strongly candidate x stands out within the local refinement window. We further measure its visual consistency with the selected anchor a_i :

$$q_i(x) = \frac{e_x^\top e_{a_i} + 1}{2}, \quad (10)$$

where e_{a_i} is the CLIP image embedding of anchor a_i . This term favors candidates that are query-relevant while remaining consistent with the local anchor context.

To suppress cross-scene drift, candidate frames are assigned to scene segments according to their temporal positions. Let $S(a_i)$ and $S(x)$ denote the scene segments containing anchor a_i and candidate frame x , respectively. We introduce a scene-evidence boundary gate:

$$b_i(x) = \eta_b + (1 - \eta_b) \chi_i(x), \quad (11)$$

where $\chi_i(x) = 1$ if x is in the same scene as a_i or belongs to an evidence-supported neighboring scene, and $\chi_i(x) = 0$ otherwise. A cross-scene candidate is regarded as evidence-supported when $E_{S(x)} \geq E_{S(a_i)} - \delta_b$, where $\delta_b = 0.02$ controls the evidence tolerance. Thus, same-scene and evidence-supported candidates are preserved, while unsupported cross-scene candidates receive a soft penalty controlled by the boundary-gate floor $\eta_b = 0.70$. Since this gate does not account for temporal distance, we further introduce a cross-scene distance decay:

$$\omega_i(x) = \exp(-\lambda_{\text{dist}} d_i(x) c_i(x)), \quad (12)$$

where $\lambda_{\text{dist}} = 0.70$ is the decay strength, t_x and t_{a_i} denote the temporal positions of candidate frame x and anchor a_i , respectively, $d_i(x) = \frac{|t_x - t_{a_i}|}{\max_{y \in W_i} |t_y - t_{a_i}| + \varepsilon}$ is the normalized temporal distance, and $c_i(x) = \mathbf{1}[S(x) \neq S(a_i)]$ indicates whether candidate x crosses a scene boundary.

Final Refinement Scoring. The final boundary-aware refinement score is:

$$r_i(x) = u_i(x) \cdot q_i(x) \cdot b_i(x) \cdot \omega_i(x). \quad (13)$$

For each selected anchor, we choose the refined frame by the refinement score, the raw query similarity, and temporal proximity to the anchor:

$$x_i^* = \arg \max_{x \in W_i} (r_i(x), s(x), -|t_x - t_{a_i}|). \quad (14)$$

The final selected frame set is obtained by collecting one refined frame for each selected anchor:

$$X^* = \{x_i^* \mid a_i \in \mathcal{A}^*\}. \quad (15)$$

By searching neighboring GOPs, this completes the selection process, converting the selected I-frame anchors into the final frame set X^* with limited neighboring-GOP decoding and scoring.

3 Experiments

3.1 Experimental Setup

Benchmarks and Backbone Models. We evaluate VISA on three long-video question answering benchmarks: VideoMME (Fu et al., 2025), MLVU (Zhou et al., 2024), and LongVideoBench (LVB) (Wu et al., 2024). All experiments on these benchmarks follow their default settings. More information about benchmarks is

shown in Appendix B. To demonstrate generalizability, we integrate VISA with three representative MLLMs: Qwen2.5-VL-7B-Instruct (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025a), and LLaVA-OneVision-Qwen2-7B-OV (Li et al., 2024a). All models are used only for inference, without training or fine-tuning.

Baselines. We compare VISA with several frame selection strategies. Base denotes the default uniform 32-frame sampling setting of each MLLM, without any additional frame selection method. For existing frame selection baselines, we compare with AKS (Tang et al., 2025), FOCUS (Zhu et al., 2025b), and WFS-SB (Chen et al., 2026). AKS selects frames by jointly considering query-frame relevance and temporal coverage. FOCUS first explores informative temporal clips to reduce the candidate search space, and then selects frames from the selected clips. WFS-SB introduces wavelet-based semantic boundary detection to obtain structure-aware candidate segments for frame selection. We use the official implementations of the baselines, follow their default settings, and re-evaluate all methods under the same MLLM backbones, benchmarks, and frame budgets. VISA uses the lightweight CLIP ViT-B/32 model (Radford et al., 2021) for anchor-level relevance estimation.

Implementation Details. We conduct all experiments using the lmms-eval framework (Zhang et al., 2025b). The inference environment is equipped with 6 NVIDIA A100 GPUs and uses PyTorch 2.5.1, CUDA 12.1, and Transformers 4.49.0. All models are evaluated under multi-GPU inference without any parameter updates. Unless otherwise specified, we keep all hyperparameters fixed across benchmarks and backbone models.

3.2 Main Results

We compare VISA with Base and recent frame selection baselines under the same 32-frame visual budget. As shown in Table 1, VISA consistently improves over Base across three MLLMs and three long-video understanding benchmarks, showing that skeleton-aware frame selection provides more informative visual evidence than uniform sampling.

VISA achieves the strongest average improvement on InternVL3-8B and Qwen2.5-VL-7B, with relative gains of +5.1% and +8.8% over Base, respectively. It also obtains the best LongVideoBench results on these two backbones,

Table 1: **Performance comparison of frame selection methods on VideoMME, MLVU, and LongVideoBench.** We report accuracy scores (%). Δ denotes the average relative improvement over Base, computed using VideoMME Overall, MLVU, and LongVideoBench. Bold indicates the best result and underline indicates the second-best result.

Model	Method	VideoMME				MLVU	LongVideoBench	Δ
		Short	Medium	Long	Overall			
LLaVA-OV-7B	Base	70.7	56.3	49.0	58.7	62.9	55.7	+0.0%
	AKS [CVPR'25]	72.0	57.9	51.2	<u>60.4</u>	65.8	<u>58.7</u>	+4.3%
	WFS-SB [CVPR'26]	73.0	63.1	50.0	62.0	<u>67.9</u>	55.8	<u>+4.6%</u>
	FOCUS [ICLR'26]	67.3	57.2	50.7	58.4	67.9	59.8	+5.0%
	VISA	70.6	59.2	50.0	59.9	66.5	58.6	+4.4%
InternVL3-8B	Base	74.1	63.2	54.2	63.9	68.6	59.2	+0.0%
	AKS [CVPR'25]	76.3	65.4	54.1	65.3	74.3	61.0	<u>+4.5%</u>
	WFS-SB [CVPR'26]	75.7	66.4	54.3	<u>65.5</u>	<u>73.6</u>	58.3	+2.8%
	FOCUS [ICLR'26]	70.9	63.9	56.0	63.6	71.8	<u>61.3</u>	+2.6%
	VISA	76.0	65.4	55.6	65.7	73.5	62.3	+5.1%
Qwen2.5-VL-7B	Base	72.8	61.8	49.7	61.4	59.3	58.6	+0.0%
	AKS [CVPR'25]	72.8	62.4	53.6	62.9	65.4	61.6	+5.9%
	WFS-SB [CVPR'26]	74.4	64.4	54.2	64.4	70.3	58.8	<u>+7.8%</u>
	FOCUS [ICLR'26]	70.2	60.9	53.4	61.5	67.9	62.3	+7.0%
	VISA	75.0	62.3	54.4	<u>63.9</u>	67.2	63.8	+8.8%

suggesting that preserving the global video skeleton is particularly useful when relevant evidence is sparsely distributed over long videos. Compared with AKS, WFS-SB, and FOCUS, VISA achieves competitive or better performance while avoiding over-reliance on locally query-relevant frames. Overall, these results demonstrate that VISA generalizes well across different MLLM backbones and long-video benchmarks, achieving strong and stable performance under the same frame budget.

3.3 Sensitivity to Frame Budget

We analyze the effect of frame budget using Qwen2.5-VL-7B. As shown in Table 2, VISA consistently improves over Base across all frame budgets on both VideoMME and LongVideoBench, demonstrating its robustness under different visual input budgets. Compared with FOCUS, VISA achieves better performance in most settings, especially on LongVideoBench, where it outperforms FOCUS under all budgets. On VideoMME, FOCUS performs slightly better at the extremely small budget of 8 frames, while VISA surpasses it as the budget increases to 16, 32, and 64 frames.

The results also show that simply increasing the number of uniformly sampled frames is not sufficient for long-video understanding. For example, Base only improves moderately when increasing the budget from 32 to 64 frames, and FOCUS even drops below Base on VideoMME at 64 frames. In contrast, VISA with 32 frames already outper-

Table 2: **Frame budget sensitivity.** Scores are reported in Acc. (%), and Δ denotes the absolute improvement over Base under the same frame budget.

Frames	Method	VideoMME		LongVideoBench	
		Acc.	Δ	Acc.	Δ
8	Base	53.5	–	52.7	–
8	FOCUS	57.2	+3.7	57.4	+4.7
8	VISA	56.6	+3.1	58.7	+6.0
16	Base	58.0	–	56.8	–
16	FOCUS	59.8	+1.8	59.9	+3.1
16	VISA	59.9	+1.9	61.6	+4.8
32	Base	61.4	–	58.6	–
32	FOCUS	61.5	+0.1	62.3	+3.7
32	VISA	63.9	+2.5	63.8	+5.2
64	Base	63.1	–	60.4	–
64	FOCUS	62.1	-1.0	62.5	+2.1
64	VISA	65.2	+2.1	64.3	+3.9

forms both Base and FOCUS with 64 frames on LongVideoBench. These results suggest that effective long-video understanding depends not only on the number of input frames, but also on selecting temporally structured and query-relevant evidence.

3.4 Efficiency Analysis

We evaluate efficiency on LongVideoBench using Qwen2.5-VL-7B under the same experimental environment and settings. On average, each video contains 20,033.7 frames. As shown in Table 3, VISA achieves the lowest decoding time and end-

Table 3: **Efficiency comparison on LongVideoBench.** “Dec.” denotes decoding, and “E2e Lat.” denotes end-to-end latency, including frame decoding, query relevance computation, frame selection, and MLLM inference.

Method	Dec. Frames	Dec. Time (s)	E2e. Lat. (s)
AKS	736.4	35.93	83.19
WFS-SB	750.9	35.74	84.03
FOCUS	348.1	36.08	58.09
VISA	408.2	22.40	39.00

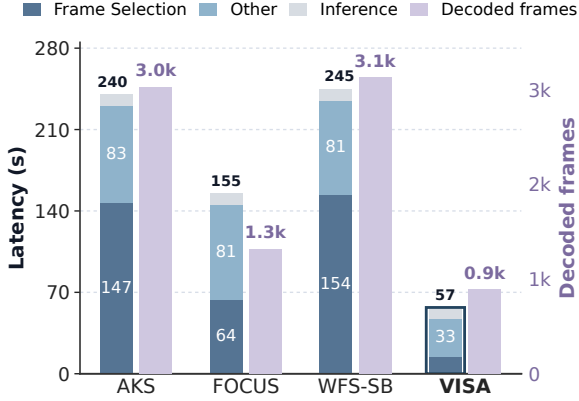


Figure 3: **Case study of end-to-end latency breakdown.** We compare decoded frames, frame selection time, MLLM inference time, and end-to-end latency on a 3024.89-second LongVideoBench video with 90,656 frames and 604 I-frames. Frame selection time includes decoding, query relevance computation, and selection. VISA achieves the lowest end-to-end latency of 57.30s. Model loading time is excluded.

to-end latency among all methods. Although VISA decodes slightly more frames than FOCUS on average, it still reduces both decoding time and overall latency. This is because VISA mainly starts from sparse I-frame anchors and only refines selected neighboring GOPs, avoiding much of the dependency decoding required by non-I frames.

We further select a representative long-video case from LongVideoBench to analyze the end-to-end latency breakdown. As shown in Fig. 3, VISA reduces both decoded frames and query relevance computation time compared with existing methods in this case. Since all methods use the same number of final selected frames, their MLLM inference times are nearly identical. Therefore, the latency reduction mainly comes from lower pre-inference costs, including frame decoding, relevance scoring, and frame selection. These results demonstrate the practical efficiency of VISA for long-video MLLM frame selection.

Table 4: **Effect of GOP-native I-frame anchors.** All variants are evaluated using Qwen2.5-VL-7B with a 32-frame budget. Scores are reported in accuracy (%), and ↓ denotes the absolute drop from GOP-native I-frame anchors.

Benchmark	Accuracy (%)		
	Uniform	Random	GOP-native
LVB	62.3 (↓1.5)	62.1 (↓1.7)	63.8
VideoMME	62.8 (↓1.1)	62.4 (↓1.5)	63.9
MLVU	63.6 (↓3.6)	64.1 (↓3.1)	67.2

3.5 Effect of GOP-native Anchors

We further isolate the effect of the initial anchor source by replacing GOP-native I-frame anchors with equally sparse non-native alternatives, including uniform sparse anchors and random sparse anchors. All variants keep the same anchor count, skeleton-aware selection, and cross-GOP refinement. As shown in Table 4, GOP-native I-frame anchors achieve the best performance, indicating that VISA benefits not only from hierarchical selection and refinement, but also from using I-frames as video-native and decoding-efficient temporal anchors for video skeleton construction.

3.6 Ablation Study

We conduct ablation studies with Qwen2.5-VL-7B to analyze the key components of VISA. As shown in Table 5, removing any component consistently degrades performance on both VideoMME and LongVideoBench.

For *Skeleton-Aware Anchor Selection*, replacing skeleton-aware allocation with global utility ranking causes the largest drop, decreasing performance by 1.1 points on VideoMME and 2.8 points on LongVideoBench. This indicates that flat anchor ranking may over-concentrate selected anchors in local segments, whereas skeleton-aware allocation better distributes the frame budget over query-relevant scenes. Removing the redundancy penalty or the scene-anchor hierarchy also leads to consistent degradation, confirming the importance of selecting locally distinctive anchors under the hierarchical video skeleton.

For *Cross-GOP Frame Refinement*, directly using selected anchors without cross-GOP refinement results in clear performance drops, suggesting that I-frame anchors provide coarse but not always optimal evidence locations. The degradation caused by removing boundary-aware scoring or restrict-

Table 5: Ablation study of VISA components.

Variant	VideoMME	LVB
Full VISA	63.9	63.8
<i>Module 2: Skeleton-Aware Anchor Selection</i>		
w/o skeleton-aware allocation	62.8 ($\downarrow 1.1$)	61.0 ($\downarrow 2.8$)
w/o redundancy penalty	62.8 ($\downarrow 1.1$)	62.5 ($\downarrow 1.3$)
w/o scene-anchor hierarchy	62.7 ($\downarrow 1.2$)	62.3 ($\downarrow 1.5$)
<i>Module 3: Cross-GOP Frame Refinement</i>		
w/o cross-GOP refinement	62.2 ($\downarrow 1.7$)	62.8 ($\downarrow 1.0$)
w/o boundary-aware scoring	63.1 ($\downarrow 0.8$)	63.5 ($\downarrow 0.3$)
single-GOP refinement	63.4 ($\downarrow 0.5$)	63.0 ($\downarrow 0.8$)

Table 6: Hyperparameter sensitivity of VISA.

Para.	Accuracy (%)				
τ	62.2 _{0.06}	63.5 _{0.08}	63.8 _{0.10}	63.6 _{0.12}	63.3 _{0.14}
η_b	62.5 _{0.50}	63.6 _{0.60}	63.8 _{0.70}	63.4 _{0.80}	63.4 _{0.90}
δ_b	63.3 _{0.005}	63.6 _{0.010}	63.8 _{0.020}	63.6 _{0.030}	63.0 _{0.040}
λ_{dist}	63.7 _{0.40}	63.7 _{0.50}	63.7 _{0.60}	63.8 _{0.70}	63.7 _{0.80}
r	63.1 ₀	63.5 ₁	63.8 ₂	63.8 ₃	63.4 ₄

ing refinement to a single GOP further shows that neighboring-GOP search and boundary-aware scoring are both useful for recovering more accurate frame-level evidence. Overall, these results verify that both skeleton-aware anchor selection and boundary-aware cross-GOP refinement are essential to VISA.

3.7 Parameter Sensitivity Analysis

We conduct parameter sensitivity analysis on Qwen2.5-VL-7B with LongVideoBench. Table 6 reports the sensitivity results of VISA to different hyperparameters. For each row, we vary one hyperparameter while keeping the others fixed at their default values. Overall, VISA remains stable across a wide range of settings, indicating that the method is not sensitive to small hyperparameter variations. The default configuration consistently achieves the best or near-best performance. In particular, λ_{dist} shows very small fluctuations, suggesting that the distance decay term mainly provides a stable refinement bias rather than requiring careful tuning. Larger deviations from the default values, such as a smaller τ or an overly large δ_b , lead to noticeable drops, showing that overly weak anchor filtering or overly loose boundary control can degrade frame selection quality.

4 Related Work

Multimodal Large Language Models for Video Understanding. Multimodal large language models (MLLMs) (Li et al., 2024a; Wang et al., 2024a; Bai et al., 2025; Xu et al., 2024; Zhang et al., 2025a) have extended visual understanding from images to videos, enabling open-ended video question answering, temporal reasoning, and multimodal instruction following. Most VideoLLMs follow a sample-encode-reason pipeline (Zhang et al., 2023; Maaz et al., 2024), where sampled frames are encoded into visual tokens and then fed into a large language model with textual instructions. However, this pipeline becomes costly for long videos, as more frames are needed to cover sparse and temporally distributed evidence, increasing the overhead of frame decoding, visual encoding, relevance estimation, and language-model inference (Song et al., 2024; Fan et al., 2024). Meanwhile, dense sampling often introduces substantial temporal redundancy and quickly consumes the limited visual context budget of MLLMs. To boost the performance of video understanding, recent studies have explored frame selection (Tang et al., 2025; Zhu et al., 2025b) and compact video representation learning (Zeng et al., 2025; Weng et al., 2024).

Frame Selection for Long Video Understanding.

Frame selection reduces redundant visual inputs by selecting frames according to query relevance, temporal coverage, semantic boundaries, or clip-level saliency (Chen et al., 2026; Hu et al., 2025; Tang et al., 2025; Zhu et al., 2025b). However, existing methods often still depend on densely decoded candidates before selection. AKS (Tang et al., 2025) and WFS-SB (Chen et al., 2026) compute relevance scores over low-rate sampled frames, which remains costly for long videos. FOCUS (Zhu et al., 2025b) improves efficiency by selecting temporal clips, but it does not explicitly maintain a global evidence structure across the video. VISA instead uses sparse GOP-native I-frame anchors to construct a video skeleton, performs skeleton-aware anchor selection, and refines only selected GOP neighborhoods, reducing decoding and scoring costs while preserving temporal coherence.

5 Conclusion

In this paper, we presented VISA, a video skeleton-aware frame selection method for efficient long-video understanding. VISA leverages GOP-native

I-frame anchors to construct a compact video skeleton, performs skeleton-aware anchor selection, and refines selected anchors within neighboring GOPs. Experiments show that VISA achieves average gains of +8.8% on Qwen2.5-VL-7B and +5.1% on InternVL3-8B over uniform sampling, while reducing end-to-end latency by 32.9%–53.6% compared with recent frame selection baselines on LongVideoBench. These results demonstrate the effectiveness of GOP-native structure for efficient and reliable long-video frame selection.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2024YDLN0004 and the Fundamental Research Funds for the Central Universities under Grant No. WK2102026004.

Limitations

Our experimental scope is currently bounded by computational resource constraints, confining our evaluation mainly to the 7B–8B parameter regimes. Although VISA demonstrates consistent efficiency and performance gains with multiple representative VideoLLMs, extending the evaluation to larger and stronger backbones would require substantially higher inference cost on long-video benchmarks. Consequently, the scaling behavior and performance ceiling of VISA with more powerful VideoLLMs remain to be fully characterized, which we leave for future work.

Ethics Statement

This work focuses on efficient long-video frame selection for MLLM-based video understanding. All datasets and benchmarks used in our experiments are publicly available and are used only for academic research and evaluation. We follow the official evaluation protocols of the corresponding benchmarks. VISA aims to reduce the computational cost of long-video understanding frame selection rather than identify individuals, track persons, or infer private attributes. Since our method relies on existing MLLMs, its outputs may still inherit model limitations such as biased predictions, hallucinations, or factual errors. Therefore, real-world deployment should carefully consider privacy, fairness, and potential misuse, especially in safety-critical or high-stakes applications. Our

experiments build upon public benchmarks, open-source models, and common research frameworks, and we respect their licenses and usage policies.

References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiaowu Zheng. 2026. Wavelet-based frame selection by detecting semantic boundary for long video understanding. *arXiv preprint arXiv:2603.00512*.
- Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer.
- Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 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*, pages 24108–24118.
- Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and 1 others. 2025. M-llm based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13702–13712.
- Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, and 1 others. 2025. Storm: Token-efficient long video understanding for multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5830–5841.
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2025. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102.
- Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024b. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.
- Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 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 (Volume 1: Long Papers)*, pages 12585–12602.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169.
- Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 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*, pages 18221–18232.
- Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and 1 others. 2024b. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pages 396–416. Springer.
- Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. 2024. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.
- Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.
- Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. 2025. Generative frame sampler for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17900–17917.
- Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, and 1 others. 2025. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pages 38057–38081.
- Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.
- Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553.
- Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jing kang Yang, Chunyuan Li, and 1 others. 2025b. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2(5):6.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025a. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

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

A Details of Skeleton-Aware Anchor Selection

Algorithm 1 describes the scene segmentation procedure used to construct the video skeleton. Starting from GOP-native I-frame anchors, we compute lightweight luma thumbnails and measure the structural change between adjacent anchors using cosine distance. The resulting change-intensity signal is smoothed with a three-point mean filter, and local peaks are detected as candidate scene boundaries. We then apply an adaptive Otsu threshold to retain prominent boundary peaks and split the I-frame skeleton into temporally coherent scenes. This process only operates on sparse GOP-native anchors, making scene segmentation efficient while preserving the major temporal structure of the video.

Algorithm 2 details the skeleton-aware anchor selection module. Given the I-frame skeleton and scene partition, we first compute query-aware anchor utilities and apply a local redundancy discount to suppress near-duplicate anchors. We then aggregate the adjusted anchor utilities into scene-level utility functions. Under the overall frame budget, a dynamic programming procedure allocates anchors across scenes, encouraging both query relevance and temporal coverage. The selected anchors are finally sorted in temporal order and sent to the boundary-aware refinement module.

B Benchmarks and Evaluation Details

To comprehensively evaluate the effectiveness of VISA, we conduct experiments on

Algorithm 1: Scene Segmentation with GOP-native Anchors

Input: GOP-native I-frame anchors $\mathcal{A} = \{a_i\}_{i=1}^M$, thumbnail size s

Output: Scene partition $\mathcal{S} = \{S_k\}_{k=1}^K$

if $M = 0$ **then**

return $\emptyset$;

if $M = 1$ **then**

return $\{[a_1, a_1]\}$;

for $i = 1$ **to** M **do**

 Decode anchor frame a_i as RGB image x_i ;

 Convert x_i to luma:

$y_i \leftarrow 0.299R_i + 0.587G_i + 0.114B_i$;

 Resize y_i to $s \times s$ and flatten it into $\bar{y}_i$;

 Normalize $\bar{y}_i \leftarrow \bar{y}_i / \|\bar{y}_i\|_2$;

for $i = 1$ **to** $M - 1$ **do**

$c_i \leftarrow \text{clip}(\bar{y}_i^\top \bar{y}_{i+1}, -1, 1)$;

$d_i \leftarrow 1 - c_i$;

$\tilde{d} \leftarrow \text{ThreePointMean}(d)$;

Initialize candidate boundary set $\mathcal{P} \leftarrow \emptyset$;

for $i = 2$ **to** $M - 2$ **do**

if $\tilde{d}_i \geq \tilde{d}_{i-1}$ **and** $\tilde{d}_i \geq \tilde{d}_{i+1}$ **then**

$\mathcal{P} \leftarrow \mathcal{P} \cup \{i\}$;

if $\mathcal{P} = \emptyset$ **then**

return $\{[a_1, a_M]\}$;

$\theta \leftarrow \text{OtsuThreshold}(\{\tilde{d}_i \mid i \in \mathcal{P}\})$;

$\mathcal{B} \leftarrow \{i \in \mathcal{P} \mid \tilde{d}_i \geq \theta\}$;

if $\mathcal{B} = \emptyset$ **then**

return $\{[a_1, a_M]\}$;

Sort $\mathcal{B}$ in ascending order;

$\mathcal{S} \leftarrow \emptyset, l \leftarrow 1$;

foreach boundary index $b \in \mathcal{B}$ **do**

$\mathcal{S} \leftarrow \mathcal{S} \cup \{[a_l, \dots, a_b]\}$;

$l \leftarrow b + 1$;

if $l \leq M$ **then**

$\mathcal{S} \leftarrow \mathcal{S} \cup \{[a_l, \dots, a_M]\}$;

return $\mathcal{S}$;

three widely used long-video understanding benchmarks: VideoMME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2024). These benchmarks cover diverse video lengths, visual domains, and reasoning tasks, allowing us to evaluate both short-term visual perception and long-range temporal reasoning. For each experimental setting, we repeat the evaluation three times and report the mean result across runs. To ensure a standardized and reproducible evaluation protocol, we conduct all experiments using lmms-eval (Zhang et al., 2025b), a unified evaluation framework for large multimodal models. The framework provides consistent task interfaces, model wrappers, and

metric computation pipelines across multimodal benchmarks, which enables fair comparison among different frame selection methods under the same evaluation setting. We provide the details of each benchmark below:

- **Video-MME**: It is a full-spectrum, multi-modal evaluation benchmark covering 6 primary visual domains and 30 subfields. It consists of 900 videos totaling 254 hours and 2,700 strictly human-annotated QA pairs. To rigorously test temporal scalability, the videos are categorized into short- (< 2 mins), medium- (4-15 mins), and long-term (30-60 mins) subsets.
- **LongVideoBench**: It is specifically designed for long-context interleaved video-language understanding. It comprises 3,763 web-collected videos (up to an hour long) paired with subtitles. The dataset includes 6,678 human-annotated multiple-choice questions spanning 17 categories. Notably, it introduces a challenging “referring reasoning” task that requires models to accurately retrieve and reason over fine-grained details within extensive video contexts.
- **MLVU**: (Multi-task Long Video Understanding) focuses on evaluating models across videos ranging from 3 minutes to 2 hours in duration. It encompasses 9 distinct evaluation tasks seamlessly categorized into holistic understanding, single detail understanding, and multi-detail understanding. The benchmark challenges models to effectively leverage both global narratives and local visual clues for complex tasks.

C Qualitative Visualization

Figure 4 presents a qualitative example from LongVideoBench using Qwen2.5-VL-7B (Bai et al., 2025) under a 32-frame budget. The question asks what the person does after introducing a blue poster, where the correct answer is “held a guitar”. We compare VISA with representative frame selection methods, including AKS (Tang et al., 2025), WFS-SB (Chen et al., 2026), and FOCUS (Zhu et al., 2025b). For each method, we visualize all 32 selected frames in an 8×4 layout, and highlight frames near the answer-relevant evidence with red boxes.

As shown in Figure 4, existing methods tend to focus on visually salient or repeatedly appearing content around the room and poster, but fail to reliably retain the temporally localized evidence that directly supports the answer. VISA, by modeling the video skeleton and refining candidate frames across GOP boundaries, better preserves the critical poster-to-guitar transition. This example demonstrates that effective long-video frame selection requires not only query relevance, but also temporally coherent coverage of answer-critical events.

Algorithm 2: Skeleton-Aware Anchor Selection

Input: I-frame anchors $\mathcal{A} = \{a_i\}_{i=1}^M$, scene partition $\mathcal{S} = \{S_k\}_{k=1}^K$, anchor relevance scores $\{p_i\}_{i=1}^M$, scene evidence $\{E_k\}_{k=1}^K$, video-level evidence E_v , anchor embeddings $\{e_i\}_{i=1}^M$, anchor budget B , redundancy radius r

Output: Selected anchor set $\mathcal{A}^*$

Define $g(m) = \log(1 + m)$ and

$w_j = g(j) - g(j - 1)$;

foreach scene $S_k \in \mathcal{S}$ **do**

$\mathcal{A}_k \leftarrow \{a_i \in \mathcal{A} \mid a_i \in S_k\}$;

$\Delta_k^{\text{scene}} \leftarrow [E_k - E_v]_+$;

foreach anchor $a_i \in \mathcal{A}_k$ **do**

$\Delta_i \leftarrow [p_i - E_k]_+$;

if $|\mathcal{N}_r(i)| > 0$ **then**

$\rho_i^{(r)} \leftarrow \frac{1}{|\mathcal{N}_r(i)|} \sum_{j \in \mathcal{N}_r(i)} \frac{e_i^\top e_j + 1}{2}$;

else

$\rho_i^{(r)} \leftarrow 0$;

$\tilde{\Delta}_i \leftarrow \Delta_i(1 - \rho_i^{(r)})$;

 Sort anchors in $\mathcal{A}_k$ by $\tilde{\Delta}_i$ in descending order;

 Denote the j -th largest effective gain as $\tilde{\Delta}_{k,(j)}$;

$U_k(0) \leftarrow 0$;

for $m = 1$ **to** $\min(B, |\mathcal{A}_k|)$ **do**

$U_k(m) \leftarrow \Delta_k^{\text{scene}} g(m) + \sum_{j=1}^m w_j \tilde{\Delta}_{k,(j)}$;

Initialize $\text{dp}[0, 0] = 0$ and $\text{dp}[0, b] = -\infty$ for $b > 0$;

Initialize backtracking table T ;

for $k = 1$ **to** K **do**

for $b = 0$ **to** B **do**

$\text{dp}[k, b] \leftarrow -\infty$;

for $t = 0$ **to** $\min(b, |\mathcal{A}_k|)$ **do**

$v \leftarrow \text{dp}[k - 1, b - t] + U_k(t)$;

if $v > \text{dp}[k, b]$ **then**

$\text{dp}[k, b] \leftarrow v$;

$T[k, b] \leftarrow t$;

$\mathcal{A}^* \leftarrow \emptyset, b \leftarrow B$;

for $k = K$ **to** 1 **do**

$m_k \leftarrow T[k, b]$;

 Add the top- m_k anchors in $\mathcal{A}_k$ ranked by $\tilde{\Delta}_i$ to $\mathcal{A}^*$;

$b \leftarrow b - m_k$;

Sort $\mathcal{A}^*$ in temporal order;

return $\mathcal{A}^*$;

Question

There is a poster hanging on the wall showing a pink semicircle positioned over a blue sea. Various cartoon characters are standing on the pink semicircle. There is a string of lights in the upper right corner of the poster, with a yellow object on the left and a red object on the right. After introducing this blue poster, what did the person with the hand do next?

A. Took out a fan
D. Held a guitar

B. Took out a book
E. Picked up a cartoon toy

C. Took out a pair of shoes

Ground Truth: D



Figure 4: Qualitative comparison of different frame selection methods on a LongVideoBench example. All methods select 32 frames. AKS, WFS-SB, and FOCUS miss the key temporal evidence required to answer the question correctly, leading to incorrect predictions. In contrast, VISA preserves the transition from the poster-introduction scene to the guitar-holding action, enabling Qwen2.5-VL-7B to predict the correct answer.