RIDGE: Region-Informed Derivative-Guided Evidence Selection for Long Video Understanding

Shanqing Xu[1] **Meng Luo**[2] **Mengchen Qian**[1] **Yuhui Gao**[1] **Siyue Peng**[1]
Xiaohan Zhong[1] **Xiaojin Zhang**[1] **Zhongyu Wei**[3] **Wei Chen**[1†] **Xiang Bai**[1]
[1]Huazhong University of Science and Technology
[2]National University of Singapore [3]Fudan University
Code: https://github.com/Xssq999/RIDGE

Abstract

Long videos contain far more visual content than Large Vision-Language Models (LVLMs) can process under a fixed visual-token budget, making frame selection essential. Existing query-aware selectors usually estimate frame-query relevance and build a compact subset from high-scoring frames. Although their mechanisms differ, the similarity sequence is still often treated primarily as values to rank or sample from, rather than as an ordered signal whose shape reflects how query-relevant evidence emerges, peaks, and fades over time. This can obscure frames that explain, contextualize, or follow an event, because such evidence may lie on the rising or falling sides of a nearby relevance peak and receive lower absolute scores. We propose RIDGE, a frame selection framework that reads the frame-query similarity curve as a temporal signal. By using local changes and curvature, RIDGE partitions the timeline into structural regions and applies region-specific selection to preserve event cores, transitions, buildup, aftermath, and contextual frames under a fixed budget. It is a lightweight post-processing step on precomputed frame-query scores and requires neither training nor iterative LVLM calls. Across four long-video benchmarks and three backbones, RIDGE achieves the best performance in most settings and remains competitive in the others.

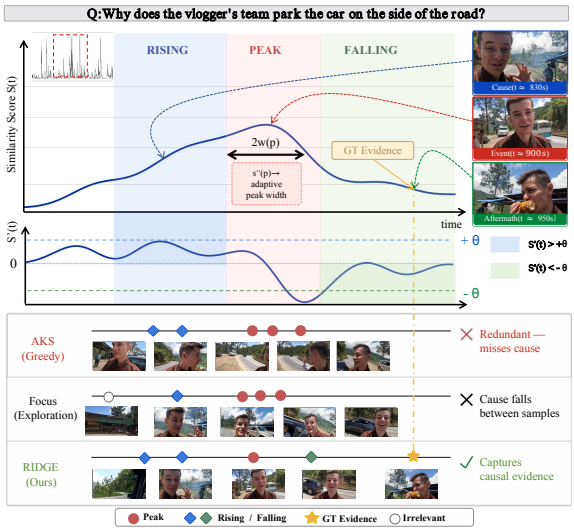

Figure 1: The question requires evidence from the cause, event, and aftermath of a long video. RIDGE samples the rising–peak–falling structure of the query-relevance curve to recover this distributed evidence.

1 Introduction

Large Vision-Language Models (LVLMs) have shown strong performance on video captioning, retrieval, and question answering (Luo et al., 2024; Maaz et al., 2024). Long-video understanding, however, remains challenging (Nguyen et al., 2024): even a one-hour video sampled at 1 FPS yields thousands of frames and millions of visual tokens. Although modern LVLMs can process increasingly long multimodal contexts, feeding such inputs in full is still computationally costly and often inefficient (Luo et al., 2026a; Tang et al., 2025a), especially when the evidence relevant to a user query lies in only a small portion of the video. Frame selection (Liang et al., 2024; Tang et al., 2025a) therefore remains a practical way to reduce visual redundancy and focus computation on query-relevant evidence. Given a video of T frames and a textual query, the selector must choose a small subset of $N \ll T$ frames that preserves the evidence needed by the downstream model (Hu et al., 2025; Luo et al., 2026b). Consequently, selecting only the highest-scoring frames can miss the buildup and aftermath evidence needed to distinguish an event from its cause or consequence.

A common query-aware approach is to compute a frame-query similarity score $s_t = \mathrm{sim}(Q, F_t)$ and select frames from this score sequence (Rad-

Email. shanqing_xu37@hust.edu.cn
†Corresponding author. lemuria_chen@hust.edu.cn

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 38717–38734
October 24-29, 2026 ©2026 Association for Computational Linguistics

ford et al., 2021; Liang et al., 2024; Hu et al., 2025). Recent methods enrich this pipeline with temporal grouping (Wang et al., 2025b), diversity objectives (Liu et al., 2025; Hu et al., 2025), adaptive search (Tang et al., 2025a; Ye et al., 2025), or iterative reasoning (Wang et al., 2025b; Ye et al., 2025). However, they still largely read the similarity curve by its magnitude: high-scoring frames are treated as the strongest evidence, while temporal structure is used mainly for coverage or search (Liu et al., 2025; Tang et al., 2025a; Ye et al., 2025). This can be limiting when the answer depends not only on the most explicit moment, but also on how that moment is reached or what changes around it.

We therefore take a different view: the frame-query similarity sequence is not merely a list of relevance scores, but a query-conditioned temporal signal. Its local shape can reveal whether the video is approaching a relevant event, staying at its core, moving away from it, crossing a sharp transition, or providing broader context. Based on this view, we propose **RIDGE**, a training-free query-aware frame selector that reads the similarity curve by both height and shape.

Given precomputed frame-query scores, RIDGE first segments the smoothed similarity curve into structural regions using local slope and curvature. It then allocates the frame budget across regions according to question preference and applies region-matched selection rules. As a result, highly relevant event cores are preserved without consuming the entire budget, leaving room for buildup, aftermath, transitions, and context that may be necessary for answering long-video questions. RIDGE is lightweight, backbone-agnostic, and does not require iterative calls to the downstream LVLM.

Our **main contributions** are threefold: 1) We reinterpret query-aware keyframe selection as reading a query-conditioned temporal signal, showing that the local shape of the similarity curve encodes different evidence roles beyond score magnitude alone; 2) We propose RIDGE, a lightweight training-free selector that segments the similarity curve into structural regions and performs question-aware, region-specific frame selection to preserve both event cores and surrounding temporal evidence; 3) We validate RIDGE on four long-video benchmarks and three LVLM backbones, demonstrating consistent gains over uniform sampling and competitive or superior performance to stronger adaptive and LVLM-assisted baselines.

2 Related Work

LVLMs for Long Video Understanding. LVLMs have rapidly evolved from image-centric multimodal systems to models capable of processing video inputs (Tang et al., 2025b; Fei et al., 2025). For long-video understanding, recent models mainly improve the way visual evidence is represented once it enters the model (Maaz et al., 2024). Some systems improve the fidelity of visual inputs through dynamic resolution, time-aware encoding, unified image-video training, or larger visual backbones, as in Qwen2.5-VL (Bai et al., 2025), LLaVA-OneVision (Li et al., 2024), and InternVL-3 (Zhu et al., 2025a). Others reduce the burden of long inputs through temporal aggregation, visual token compression, or memory-style mechanisms, as in Video-LLaMA (Zhang et al., 2023), LLaVA-Video (Zhang et al., 2024b), and LongVU (Shen et al., 2024). These advances strengthen the downstream reasoner (Shu et al., 2025; Zhang et al., 2024a), but they still operate on a finite visual input. In practical long-video QA, the model usually sees a sampled or compressed view of the original video (Song et al., 2024). Architectural scaling can improve how the model uses visible evidence, while frame selection determines which evidence becomes visible before reasoning begins.

Keyframe Selection. Keyframe selection controls video input view by choosing a compact set of frames for downstream reasoning. Query-agnostic strategies, such as uniform sampling or shot-based summarization, provide broad temporal coverage but cannot adapt to the evidence required by a particular question. Query-aware methods therefore build a relevance signal between the query and candidate frames or clips. The simplest form is top-k selection, while stronger variants use the same signal together with temporal structure or selection constraints. AKS (Tang et al., 2025a), for example, balances relevance and coverage through recursive temporal partitioning; MDP^3 (Sun et al., 2025) adds diversity so that selected frames are less redundant; and stochastic selectors such as BOLT (Liu et al., 2025) and Q-Frame (Zhang et al., 2025b) reduce the brittleness of always choosing the highest-scoring frames. Another family allocates computation more actively: FOCUS (Zhu et al., 2025b) searches over clips under a fixed budget, while A.I.R. (Zou et al., 2025) and Frame-Voyager (Yu et al., 2025) use LVLM feedback or

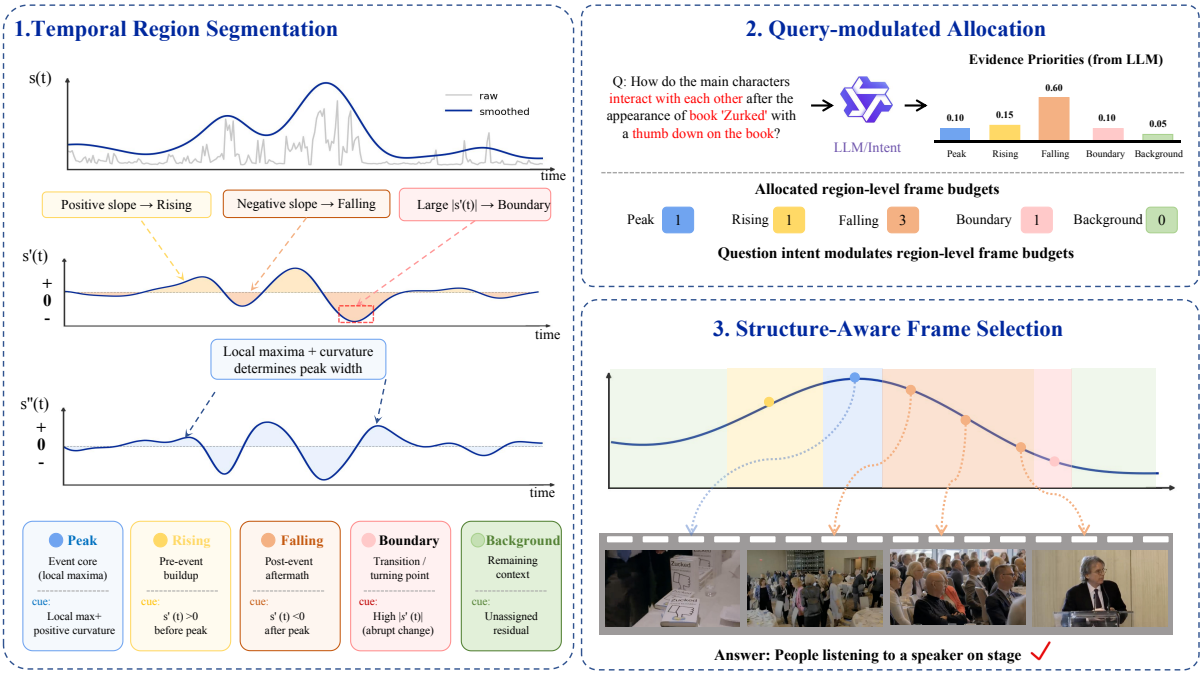

Figure 2: Overview of RIDGE. Given query-frame relevance scores, RIDGE segments the temporal signal into evidence roles, allocates the frame budget with question-aware preferences, and selects role-specific frames for LVLM inference.

supervision to refine the candidate evidence. These methods improve keyframe selection from different angles: coverage, diversity, stochasticity, adaptive search, or model-guided refinement. From the perspective of RIDGE, what remains underexplored is the internal temporal shape of the relevance signal.

3 Method

3.1 Overview

Let $V = \{F_1, \ldots, F_T\}$ be the candidate frames of a video and Q be the question. A lightweight vision-language scorer produces a query-frame relevance sequence

$$s_t = \mathrm{sim}(Q, F_t), \quad t = 1, \ldots, T.$$

Given a frame budget N, RIDGE selects $\min(N, T)$ frame indices and passes the corresponding frames to the downstream LVLM in temporal order. RIDGE also uses a compact evidence-preference vector

$$\mathbf{w} = (w_{\text{peak}}, w_{\text{slope}}, w_{\text{rise}}, w_{\text{fall}}, w_{\text{boundary}}, w_{\text{context}}),$$

which is either fixed or generated once from the question.

In many video questions, the answer is not confined to a single instant: the lead-up to an event, its immediate aftermath, or the moment where relevance changes abruptly may be just as informative as the highest-scoring frame itself. Our method RIDGE is built around a simple premise: ***a relevance curve should not be read only by its height***. Its shape also tells us whether the video is moving toward a relevant event, staying at the event core, moving away from it, or crossing an abrupt transition.

This suggests that frame selection should answer three distinct questions: 1) ***what temporal role each frame plays***; 2) ***how much budget each role deserves for the current question*** and 3) ***which representative frames best express that role***. RIDGE therefore separates frame selection into three decisions. First, it converts the one-dimensional relevance curve into a temporal role map with five regions:

$$\mathcal{Y} = \{\textsc{peak}, \textsc{rising}, \textsc{falling}, \textsc{boundary}, \textsc{background}\}.$$

PEAK frames capture event cores, RISING frames capture pre-event buildup, FALLING frames capture aftermath, BOUNDARY frames capture abrupt changes, and BACKGROUND frames preserve remaining context. Second, RIDGE allocates the frame budget across these roles according to the evidence needed by the question. Third, it selects

frames inside each region with a rule matched to that region's temporal function.

3.2 Temporal Region Segmentation

RIDGE starts by making the score sequence stable enough for temporal analysis. It min–max normalizes s_t when the score range is non-zero, smooths the normalized sequence with a Gaussian kernel, and computes its first- and second-order differences:

$$\bar{s}_t = (G_\sigma * \hat{s})(t), \qquad s'_t = \nabla \bar{s}_t, \qquad s''_t = \nabla s'_t.$$

Algorithm 1 RIDGE: Curve-guided and Query-aware Frame Selection

Require: Scores $s_{1:T}$, budget N, preference vector $\mathbf{w}$, hyperparameters Θ
Ensure: Selected indices $\mathcal{I}$ in temporal order
if $T \leq N$ **then**
  return $\{1, \ldots, T\}$
end if
Normalize and smooth scores; compute $s' = \nabla \bar{s}$ and $s'' = \nabla s'$
Compute coverage intent c from $\mathbf{w}$; initialize labels as BACKGROUND

Stage 1: curve-to-region segmentation
Set $\mathcal{P}$ to peaks on $\bar{s}$; if empty, use $\{\arg\max_t \bar{s}_t\}$
for each $p \in \mathcal{P}$ **do**
  Label a curvature-adaptive window around p as PEAK
  Extend to adjacent positive/negative slope spans as RISING/FALLING, modulated by c
end for
Label residual high-$|s'|$ frames as BOUNDARY; keep the rest as BACKGROUND

Stage 2: question-aware budget allocation
Map $\mathbf{w}$ to region priorities W_r
Allocate region budgets n_r proportionally, with capacity caps and leftover redistribution

Stage 3: region-specific frame selection
$\mathcal{I} \leftarrow \emptyset$
for each region r with $n_r > 0$ **do**
  if $r \in \{\text{RISING}, \text{FALLING}\}$ and $n_r \geq 3$ and $|\mathcal{R}_r| \geq 3$ **then**
    $\mathcal{I} \leftarrow \mathcal{I} \cup \text{BinSelect}(\mathcal{R}_r, \bar{s}, n_r)$
  else if $r = \text{BOUNDARY}$ **then**
    $\mathcal{I} \leftarrow \mathcal{I} \cup \text{RankSelect}(\mathcal{R}_r, |s'|, n_r)$
  else
    $\mathcal{I} \leftarrow \mathcal{I} \cup \text{RankSelect}(\mathcal{R}_r, \bar{s}, n_r)$
  end if
end for
Deduplicate $\mathcal{I}$; if $|\mathcal{I}| < N$, backfill unselected frames by descending $\bar{s}_t$
return first $\min(N, T)$ temporally sorted indices in $\mathcal{I}$

The smoothed score $\bar{s}_t$ estimates how relevant a frame is, the slope s'_t describes whether relevance is increasing or decreasing, and the curvature s''_t helps determine whether a relevance peak is sharp or temporally extended. This turns the score sequence from isolated framewise values into a local temporal profile whose level, slope, and sharpness can be read as complementary evidence cues.

Event cores. RIDGE first finds local maxima on $\bar{s}$ and treats them as candidate event cores. Around each peak p, it assigns a curvature-adaptive PEAK window. The half-width is inversely related to the relative curvature

$$r_p = \frac{|s''_p|}{\text{median}(|s''|) + \epsilon}.$$

RIDGE retains all local maxima that pass the prominence threshold; the global maximum is used only as a fallback when no peak is detected. Each detected peak receives its own PEAK window and adjacent RISING/FALLING extensions, although RIDGE does not impose a hard frame quota for every peak. Appendix G provides a quantitative analysis of this multi-peak behavior. Sharp peaks receive compact windows because the relevant evidence is temporally concentrated; flatter peaks receive wider windows because the event remains visually relevant for longer. This fallback ensures that every video has at least one event core, even in weakly structured or noisy cases, preventing the method from degenerating into pure background selection.

Buildup and aftermath. After locating event cores, RIDGE searches outward from each peak to recover neighboring temporal phases. Frames on the left are labeled RISING while the slope remains sufficiently positive, and frames on the right are labeled FALLING while the slope remains sufficiently negative.

The extension is modulated by the question. Given the evidence preference vector $\mathbf{w}$, RIDGE computes a coverage intent

$$c = \frac{w_{\text{rise}} + w_{\text{fall}} + w_{\text{context}}}{w_{\text{peak}} + w_{\text{rise}} + w_{\text{fall}} + w_{\text{context}} + \epsilon}.$$

A larger c means the question is expected to require broader temporal evidence, so the rising/falling search uses a looser slope threshold and a longer maximum extension. A smaller c keeps the regions tighter around the event core. Intuitively, some questions behave like snapshot queries, while others can only be answered by tracing a short temporal development. The exact thresholding rule is listed in Appendix A.

Transitions and context. Once peak, rising, and falling regions have been assigned, RIDGE looks only at the residual background frames. Residual frames with unusually large $|s'_t|$ are labeled as BOUNDARY, capturing abrupt changes in the query-relevance trajectory that may correspond to scene cuts or action switches. All remaining frames retain the BACKGROUND label.

3.3 Question-aware Budget Allocation

The region map identifies the temporal role of each frame, while $\mathbf{w}$ specifies which evidence types are likely to be useful for the question. A question may focus mainly on the event core or require a broader account of how the event is prepared, interrupted, or resolved. The entries of $\mathbf{w}$ are not frame scores; they describe what kinds of visual evidence are likely to support the answer.

RIDGE maps the six evidence preferences to five region priorities by letting PEAK, RISING, FALLING, and BACKGROUND inherit their corresponding weights, while setting $W_{\text{BOUNDARY}} = w_{\text{boundary}} + w_{\text{slope}}$. Here, w_{rise} and w_{fall} encode directional phase preferences, whereas w_{slope} encodes rapid relevance change regardless of direction. We route w_{slope} to BOUNDARY because BOUNDARY contains residual high-$|s'|$ frames and is the only region whose internal selection rule ranks by $|s'|$. Routing it additionally to RISING and FALLING could double-count slope preference in phases that already have dedicated weights.

For each non-empty region, RIDGE assigns a proportional share of the frame budget according to W_r, caps the share by the number of available frames in that region, and redistributes any unused budget to the remaining available regions. This keeps the allocation faithful to the question while respecting the actual structure discovered in the video.

3.4 Role-specific Frame Selection

The final stage chooses concrete frames from each allocated region. RIDGE uses score ranking where high absolute relevance is the right criterion, and temporal coverage where the region's value lies in its evolution over time.

For PEAK, RIDGE ranks frames by $\bar{s}_t$, keeping the clearest views of the event core. BACKGROUND uses the same ranking so that contextual frames remain query-relevant. For BOUNDARY, RIDGE ranks by $|s'_t|$, because what matters there is not how relevant a frame already is, but how strongly it marks a temporal change in relevance. For RISING and FALLING, RIDGE uses temporal sub-bin selection when the region and its budget are large enough: it splits the region into equal temporal bins and selects the highest-$\bar{s}_t$ frame from each bin. The goal is not only to keep relevant frames, but to preserve the temporal spread of evidence inside a phase. This avoids collapsing all buildup or aftermath frames near the adjacent peak. When a region is too small for meaningful binning, RIDGE falls back to score ranking.

Frames selected from all regions are merged, deduplicated, and sorted by time. If the selected set is still smaller than the target budget, RIDGE backfills from the remaining frames by descending $\bar{s}_t$. This preserves the role-aware structure whenever possible, while still using the available budget on frames that remain broadly relevant to the question.

4 Experiments

4.1 Experimental Setup

Datasets and Baselines. We evaluate RIDGE on four widely used long-video understanding benchmarks. Video-MME (Fu et al., 2025) consists of 900 videos covering short, medium, and long durations and is annotated with 2,700 multiple-choice QA pairs. We report results without subtitles. MLVU (Zhou et al., 2025) contains 1,730 long videos with an average duration of 12 minutes and covers nine task types. We report M-Avg on the dev set. LongVideoBench (LVB) (Wu et al., 2024) provides 6,678 QA pairs over 3,763 videos ranging from 8 seconds up to 1 hour, and emphasizes long-context referring reasoning. We report results on the validation set. LVBench (Wang et al., 2025a) contains 103 hour-long videos and 1,549 QA pairs requiring long-horizon temporal reasoning.

We compare RIDGE with uniform sampling, score-ranking or adaptive selectors including AKS (Tang et al., 2025a), BOLT (Liu et al., 2025), and FOCUS (Zhu et al., 2025b), and the diversity-aware selector MDP^3 (Sun et al., 2025). All baseline and competing-method results in Table 1 were obtained by rerunning the official implementations under our unified evaluation pipeline, using each method's default configuration. The main comparison retains each method's default scorer: AKS and FOCUS use BLIP-ITM, BOLT uses CLIP, MDP^3 uses SigLIP, and RIDGE uses BLIP-ITM; therefore, Table 1 is not a shared-scorer comparison. Table 2 reports matched-scorer controls that isolate

Table 1: Comparison of LVLMs and frame selection methods on Video-MME, LongVideoBench, MLVU, and LVBench. We report accuracy scores (%). **Bold** and underline indicate the best and second-best performance within each model group.

Model	Method	Frame	Video-MME	LongVideoBench	MLVU	LVBench
LLaVA-OV-7B	Baseline	32	56.5	54.8	63.1	40.4
	AKS	32	58.4 ↑1.9	59.3 ↑4.5	67.5 ↑4.4	47.3 ↑6.9
	MDP^3	32	58.5 ↑2.0	59.0 ↑4.2	**68.2** ↑5.1	41.6 ↑1.2
	BOLT	32	57.9 ↑1.4	59.6 ↑4.8	66.8 ↑3.7	38.2 ↓2.2
	FOCUS	32	58.3 ↑1.8	**60.7** ↑5.9	67.1 ↑4.0	47.5 ↑7.1
	RIDGE	32	**59.6** ↑3.1	60.1 ↑5.3	68.0 ↑4.9	**48.9** ↑8.5
InternVL-3-8B	Baseline	32	65.6	58.5	68.4	43.5
	AKS	32	66.3 ↑0.7	61.5 ↑3.0	74.2 ↑5.8	52.4 ↑8.9
	MDP^3	32	**66.8** ↑1.2	60.9 ↑2.4	73.9 ↑5.5	47.3 ↑3.8
	BOLT	32	66.1 ↑0.5	59.2 ↑0.7	70.8 ↑2.4	44.2 ↑0.7
	FOCUS	32	64.3 ↓1.3	61.0 ↑2.5	72.0 ↑3.6	53.2 ↑9.7
	RIDGE	32	66.6 ↑1.0	**61.8** ↑3.3	**74.6** ↑6.2	**54.5** ↑11.0
Qwen2.5-VL-7B	Baseline	32	61.2	58.9	59.7	38.5
	AKS	32	63.1 ↑1.9	63.2 ↑4.3	67.2 ↑7.5	43.4 ↑4.9
	MDP^3	32	**63.8** ↑2.6	60.0 ↑1.1	66.2 ↑6.5	41.3 ↑2.8
	BOLT	32	63.2 ↑2.0	60.0 ↑1.1	68.4 ↑8.7	39.3 ↑0.8
	FOCUS	32	63.4 ↑2.2	61.0 ↑2.1	68.2 ↑8.5	46.3 ↑7.8
	RIDGE	32	63.7 ↑2.5	**65.4** ↑6.5	**69.6** ↑9.9	**50.9** ↑12.4

the selection policy.

Implementation Details. We evaluate RIDGE on three representative LVLM backbones: Qwen2.5-VL-7B (Bai et al., 2025), InternVL-3-8B (Zhu et al., 2025a), and LLaVA-OV-7B (Li et al., 2024), feeding the selected frames in temporal order and following the official inference settings of each backbone. Unless otherwise stated, query–frame similarities $\{s_t\}$ are computed with BLIP-ITM (Li et al., 2022). RIDGE uses a 6-dimensional evidence-preference vector $\mathbf{w}$, instantiated either as a fixed default vector or as question-specific weights generated once by Qwen3-8B (Yang et al., 2025). We ablate Qwen3-3B and Llama-3.1-8B (Grattafiori et al., 2024) as alternative weight generators, and CLIP (Radford et al., 2021), BLIP-2 (Li et al., 2023), SigLIP (Zhai et al., 2023), and BLIP-ITM (Li et al., 2022) as scoring models. The default frame budget is N=32. RIDGE's frame-selection stage is deterministic. We use one shared RIDGE hyperparameter configuration across benchmarks, LVLM backbones, frame budgets, model scales, and scorers; the sensitivity sweeps were conducted after fixing the default and are used only as robustness analyses. RIDGE itself is a training-free, plug-in selector that runs as a lightweight CPU post-processing step and incurs no extra cost during LVLM inference. In our experiments, LVLM inference is performed on four NVIDIA A100 GPUs, and the pipeline is also deployed and applied on Ascend graphics cards. We report accuracy measured by the LMMs-Eval toolkit (Zhang et al., 2025a). Full hyperparameter settings are deferred to the appendix.

4.2 Comparison with SOTA Methods

Table 1 reports our main comparison on Video-MME, LVB, MLVU, and LVBench across three LVLM backbones. RIDGE consistently improves over uniform sampling at the same frame budget and remains competitive with stronger adaptive and diversity-aware selectors. On Qwen2.5-VL-7B, RIDGE improves the uniform baseline by 2.5 percentage points on Video-MME, 6.5 on LVB, 9.9 on MLVU, and 12.4 on LVBench. The same pattern holds for InternVL-3-8B and LLaVA-OV-7B, showing that the gains are not tied to one downstream LVLM. The improvements are especially large on MLVU and LVBench, which contain more long-horizon and multi-event reasoning cases, supporting our hypothesis that temporal-shape-aware selection is particularly beneficial when evidence is distributed across event buildup, core moments, and aftermath. The comparatively modest gain on Video-MME is expected: its benchmark is dominated by short- and medium-length clips whose key

information is often concentrated in a few visually distinct shots, leaving limited room for a smarter selection policy to add value.

Matched-Scorer Comparison. The main comparison in Table 1 retains the scorer used by each method's default configuration. To separate scorer choice from the frame-selection policy, Table 2 compares RIDGE with BOLT and MDP^3 under their respective scorers. All results use Qwen2.5-VL-7B with a frame budget of N=32.

With CLIP scores, RIDGE improves the four-benchmark average over BOLT by 3.5 percentage points. With SigLIP scores, it improves the average over MDP^3 by 3.8 percentage points. RIDGE is higher in seven of the eight matched comparisons; the exception is Video-MME with SigLIP, where it is 1.3 percentage points below MDP^3. These controls show that the gains in Table 1 are not attributable to scorer choice alone.

Effect of Question-aware Weighting. To isolate the contribution of question-specific weight generation, we compare the default Qwen3-8B generator with Qwen3-0.6B and a fixed weight vector shared by all questions. Table 3 also includes the LLM-free baselines FOCUS and AKS for reference. All results use Qwen2.5-VL-7B with a frame budget of N=32.

RIDGE with fixed weights achieves a four-benchmark average of 61.1, outperforming the LLM-free baselines FOCUS and AKS by 1.4 and 1.9 percentage points, respectively. This result shows that temporal region modeling and region-specific selection remain effective without an auxiliary weight-generation LLM. Question-specific weights generated by Qwen3-8B further improve the average to 62.4, while Qwen3-0.6B retains most of this benefit with a substantially smaller auxiliary model.

4.3 Robustness Analysis

We stress-test RIDGE along five axes: the frame budget, the scale of the downstream LVLM, the choice of frame–query scoring model, the LLM used for weight generation, and the smoothing scale σ. Unless otherwise stated, experiments use LVB with Qwen2.5-VL-7B as the default backbone and BLIP-ITM as the default scoring model; when the LLM weight mode is used, Qwen3-8B is the default weight generator.

Robustness to Frame Budget. Figure 3 reports accuracy on LVB at $N \in \{8, 16, 32, 64\}$. RIDGE improves over uniform sampling across all budgets.

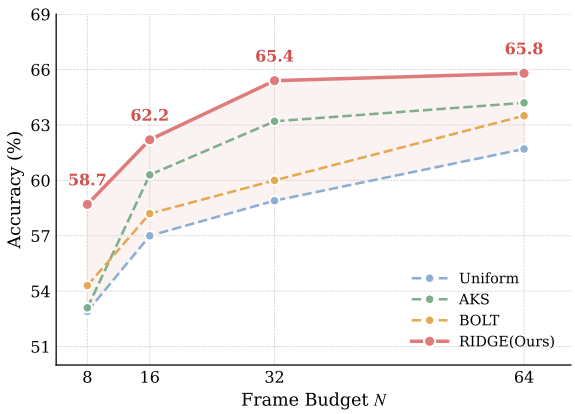

Figure 3: Accuracy on LongVideoBench under different frame budgets.

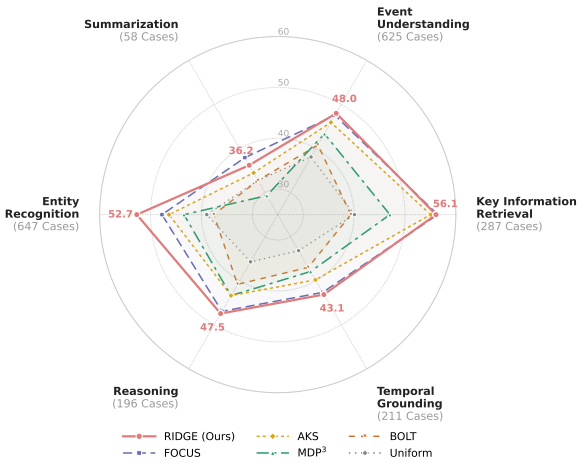

Figure 4: Per-question-type accuracy on LVBench.

The gain is positive even at N=64, indicating that the method is not only useful under extremely tight frame budgets. The margin over AKS is largest at smaller budgets, consistent with the intuition that explicit modeling of rising and falling regions matters most when redundant peak-centered frames are costly.

Robustness to LVLM Scale. We study whether the benefit of RIDGE persists as the downstream LVLM scales up. Table 4 compares uniform sampling and RIDGE on Qwen2.5-VL-3B/7B/32B/72B at N=32. RIDGE improves performance at every scale. While the absolute gain remains relatively stable across model sizes, even the 72B model still benefits by 6.3 percentage points, indicating that better frame selection is complementary to LVLM scaling rather than absorbed by larger models.

Robustness to Frame Scoring Model. RIDGE operates on a precomputed similarity sequence

Table 2: Matched-scorer comparison under Qwen2.5-VL-7B with N=32. Accuracy (%) is reported. Within each scorer block, the baseline and RIDGE use the same frame–query scorer.

Scorer	Method	Video-MME	LVB	MLVU	LVBench	Avg.
CLIP	BOLT	63.2	60.0	68.4	39.3	57.7
	RIDGE	64.3	62.0	69.7	48.9	61.2
SigLIP	MDP^3	63.8	60.0	66.2	41.3	57.8
	RIDGE	62.5	64.7	71.2	47.8	61.6

Table 3: Effect of question-aware weight generation across four benchmarks. Accuracy (%) is reported. Avg. is the average over the four benchmarks, and Δ is measured relative to Qwen3-8B.

Weight Source / Method	Video-MME	LVB	MLVU	LVBench	Avg.	Δ
Qwen3-8B (default)	63.7	65.4	69.6	50.9	62.4	0.0
Qwen3-0.6B	63.2	64.9	69.9	49.7	61.9	-0.5
RIDGE (fixed weights)	62.9	63.8	69.4	48.3	61.1	-1.3
FOCUS	63.4	61.0	68.2	46.3	59.7	-2.7
AKS	63.1	63.2	67.2	43.4	59.2	-3.2

Table 4: Robustness across LVLM scales on LongVideoBench.

Model Scale	Uniform	+RIDGE	Δ
Qwen2.5-VL-3B	54.7	59.3	+4.6
Qwen2.5-VL-7B	58.9	65.4	+6.5
Qwen2.5-VL-32B	59.5	65.5	+6.0
Qwen2.5-VL-72B	60.6	66.9	+6.3

Table 5: Sensitivity to the frame–query scoring model across benchmarks. We select BLIP-ITM as the default.

Scoring Model	LVB	Video-MME	LVBench
Uniform	58.9	61.2	38.5
CLIP	62.0	64.3	48.9
BLIP-2	63.4	63.4	49.7
SigLIP	64.7	62.5	47.8
BLIP-ITM	65.4	63.7	50.9

Table 6: Sensitivity to the weight-generation LLM on LongVideoBench.

Weight Source	LVB Acc.	Δ vs Qwen3-8B
Qwen3-8B (default)	65.4	—
Qwen3-3B	64.2	-1.2
Llama-3.1-8B	64.5	-0.9

$\{s_t\}$ and is therefore decoupled from the choice of frame scoring model. Table 5 compares different scoring backbones. Across LVB, Video-MME, and LVBench, RIDGE consistently improves over uniform sampling under different scoring models. Based on its stable performance, we use BLIP-ITM as the default scoring model in our main experiments.

Robustness to Weight-Generation LLM. We test how sensitive RIDGE is to the LLM used for generating the region weights in Table 6. Across Qwen3-3B, Qwen3-8B, and Llama-3.1-8B, the final accuracy changes moderately but remains well above the uniform baseline. This stability is expected: the LLM affects only a compact evidence-preference vector, which modulates the rising/falling extension through coverage intent and then redistributes a fixed frame budget across structural regions.

Robustness to Smoothing Scale σ. The Gaussian smoothing scale σ controls the granularity of temporal structure recovered from $\{s_t\}$. A small value leaves noise in the raw similarity curve, while a large value may merge adjacent events. Table 7 sweeps $\sigma \in \{1.0, 2.0, 3.0, 5.0, 8.0\}$. Accuracy peaks at σ=2.0 and degrades gracefully when the curve is either under-smoothed or over-smoothed, indicating that RIDGE is not overly sensitive to this hyperparameter.

Across these axes, RIDGE consistently remains above the corresponding uniform baseline. The gains are therefore not an artifact of a single frame budget, LVLM, scoring model, weight generator, or smoothing setting.

4.4 Ablation and Analysis

While Section 4.3 shows that RIDGE is robust to external design choices, this subsection dissects the internal contribution of each component in the final algorithm.

Component Ablation. Table 8 ablates RIDGE at the module level rather than listing low-level operators. We remove or replace one module from

Table 7: Sensitivity to the smoothing scale σ on LongVideoBench.

σ	LVB Acc.	Δ vs Uniform
1.0	63.0	+4.1
2.0	65.4	+6.5
3.0	63.9	+5.0
5.0	64.0	+5.1
8.0	62.8	+3.9

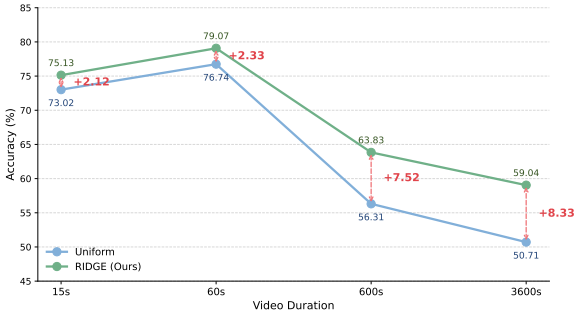

Figure 5: Accuracy on LongVideoBench grouped by video duration.

the full method at a time: temporal phase modeling covers pre-event and post-event regions; transition modeling covers abrupt-change evidence; adaptive event-core localization controls event-core spans using local curve geometry; question-aware evidence weighting controls the preference vector; and question-adaptive phase extension uses this preference to modulate how far rising and falling regions extend.

Question-type Analysis. A core motivation of RIDGE is to recover evidence distributed across different temporal roles. Figure 4 breaks down LVBench by question type. RIDGE improves substantially on key information retrieval, entity recognition, reasoning, and temporal grounding, all of which often require locating sparse evidence rather than only recognizing the most salient event frame. The gains are smaller on summarization, where broad coverage is helpful but the answer is less tied to a single rising–peak–falling structure.

Video-duration Analysis. We further analyze RIDGE on LongVideoBench by grouping videos according to their duration, as shown in Figure 5. RIDGE consistently outperforms uniform sampling across all duration buckets. The gains are moderate for short videos, with accuracy improving from 73.02 to 75.13 on the 15-second bucket and from 76.74 to 79.07 on the 60-second bucket. As videos become longer, the advantage increases substantially, reaching +7.52 percentage points on the 600-second bucket and +8.33 percentage points on the 3600-second bucket. This trend supports the motivation of RIDGE: uniform sampling can already cover many salient moments in short clips, while longer videos require preserving more temporally distributed evidence under the same frame budget.

Table 8: Component ablation on LongVideoBench. Each row removes or replaces one component from the full RIDGE pipeline. "w/o Question-aware evidence weighting" uses one fixed weight vector for all questions, whereas "w/ Uniform evidence allocation" assigns equal weights across evidence types.

Configuration	Acc.	Δ
Uniform sampling	58.9	–
RIDGE (full)	**65.4**	–
Region modeling		
w/o Temporal phase modeling	64.3	-1.1
w/o Transition modeling	64.2	-1.2
w/o Adaptive event-core localization	64.3	-1.1
Budget allocation		
w/o Question-aware evidence weighting	63.8	-1.6
w/o Question-adaptive phase extension	64.0	-1.4
w/ Uniform evidence allocation	64.4	-1.0
Selection policy		
w/o Relevance-based ranking	64.3	-1.1
w/o Temporal bin selection	64.8	-0.6

5 Conclusion

We presented RIDGE, a training-free frame selection framework for long-video understanding. The central idea is to treat the frame-query similarity sequence as an ordered temporal signal rather than a flat relevance list. By segmenting this signal into structural regions, allocating budget according to question-dependent evidence needs, and selecting frames with region-specific rules, RIDGE preserves not only salient event cores but also the buildup, aftermath, transitions, and contextual evidence that are often suppressed by score ranking. Across four long-video benchmarks and three LVLM backbones, RIDGE delivers consistent gains over uniform sampling and strong keyframe-selection baselines, with especially clear benefits in settings that require sparse or temporally distributed evidence. These results suggest that improving long-video QA is not only a matter of scoring frames more accurately.

Limitations

RIDGE is most useful when the answer depends on evidence distributed across multiple temporal

roles. For short videos, or for questions that can be answered from a single salient frame, the advantage over uniform sampling or simple score ranking may be modest because there is less temporal structure to recover. The method also depends on the quality of the frame-query scorer and the downstream LVLM. If the scorer does not produce meaningful variation around the relevant evidence, or if the LVLM cannot interpret the selected visual evidence, better frame allocation alone cannot fully solve the question. Finally, RIDGE selects from a predefined candidate frame sequence. Very brief events, fine-grained OCR, small objects, or evidence missed by the initial candidate sampling can remain inaccessible. Future work could combine temporal-shape-aware selection with denser candidate generation, stronger frame scorers, or models that jointly refine scoring and reasoning.

Acknowledgments

This research was supported by the *National Natural Science Foundation of China* (No. 62406121) and the *Natural Science Foundation of Hubei Province, China* (No. 2024AFB189).

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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. *Preprint*, arXiv:2502.13923.

Yulin Fei, Yuhui Gao, Xingyuan Xian, Xiaojin Zhang, Tao Wu, and Wei Chen. 2025. Do current video llms have strong ocr abilities? a preliminary study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9860–9876.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, 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*, pages 24108–24118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. 2025. M-llm based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13702–13712.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. 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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. 2024. Keyvideollm: Towards large-scale video keyframe selection. *arXiv preprint arXiv:2407.03104*.

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

Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. 2024. Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7667–7676.

Meng Luo, Bobo Li, Shanqing Xu, Shize Zhang, Qiuchan Chen, Menglu Han, Wenhao Chen, Yanxiang Huang, Hao Fei, Mong-Li Lee, et al. 2026a. Unveiling the cognitive compass: Theory-of-mind-guided multimodal emotion reasoning. *arXiv preprint arXiv:2602.00971*.

Meng Luo, Shengqiong Wu, Liqiang Jing, Tianjie Ju, Li Zheng, Jinxiang Lai, Tianlong Wu, Xinya Du, Jian Li, Siyuan Yan, et al. 2026b. Dr. v: A hierarchical perception-temporal-cognition framework to diagnose video hallucination by fine-grained spatial-temporal grounding. *International Journal of Computer Vision*, 134(6):278.

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.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See Kiong Ng, and Luu Anh Tuan. 2024. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3636–3657.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

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, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, 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*, pages 18221–18232.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. 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*, pages 24090–24101.

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

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2025b. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. 2025a. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025b. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, 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*, pages 8579–8591.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. 2025. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, volume 2025, pages 84154–84179.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

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, Jingkang Yang, Chunyuan Li, et al. 2025a. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. 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*, pages 22056–22065.

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

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025a. Internvl3: 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*.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. 2025. Air: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. *arXiv preprint arXiv:2510.04428*.

A Detailed RIDGE Procedure

This section expands the deterministic implementation details omitted from the main method section. The goal is to make the exact region definitions and allocation rules reproducible while keeping the main paper focused on the motivation and high-level design.

Signal preprocessing. Given raw similarity scores $\{s_t\}_{t=1}^{T}$, we first apply min–max normalization,

$$\hat{s}_t = \frac{s_t - s_{\min}}{s_{\max} - s_{\min}},$$

with the degenerate constant-score case handled by leaving the sequence unchanged. The normalized signal is smoothed with a one-dimensional Gaussian kernel,

$$\bar{s}_t = (G_\sigma * \hat{s})(t),$$

and the first- and second-order temporal differences are computed as $s'_t = \nabla \bar{s}_t$ and $s''_t = \nabla s'_t$.

Adaptive peak width. Peaks are detected on $\bar{s}_t$ with prominence threshold $p_{\min}$ and minimum peak distance of 5. If no peak is detected, the global maximum is used. For each peak p, we compute its relative curvature

$$r_p = \frac{|s''_p|}{\text{median}(|s''|) + \epsilon}.$$

The final implementation treats PEAK as the event core rather than the whole transition neighborhood. We therefore use an inverse curvature mapping: sharp peaks receive compact core windows, while flatter peaks receive wider windows. Specifically, we set

$$h_p = \text{clip}\left(\text{round}\left(\frac{w_0}{\sqrt{\max(r_p, 0.25)}}\right),\ 2,\ w_{\max}\right).$$

This keeps sharp temporal changes around narrow peaks available to the RISING, FALLING, or BOUNDARY labels.

Coverage-modulated region thresholds. From the question-preference vector, we compute

$$c = \frac{w_{\text{rise}} + w_{\text{fall}} + w_{\text{context}}}{w_{\text{peak}} + w_{\text{rise}} + w_{\text{fall}} + w_{\text{context}} + \epsilon}.$$

The rising/falling slope threshold is then

$$\tau_s = (k_s(1 - c) + 0.1c) \cdot \text{std}(s'),$$

and the maximum one-sided extension is

$$L_{\max} = \min(\text{round}(k_e \max(\sigma, 1)(1 + c)),\ \lfloor T/4 \rfloor).$$

Starting from each peak boundary, we extend left while $s'_t > \tau_s$ and right while $s'_t < -\tau_s$, up to $L_{\max}$. For boundary detection, we only consider frames that are still labeled as background after peak and slope-region assignment. Let this residual set be $\mathcal{B}$. A frame $t \in \mathcal{B}$ is labeled as boundary if

$$|s'_t| > \text{median}(|s'|_{\mathcal{B}}) + k_b \cdot \text{std}(|s'|_{\mathcal{B}}).$$

Question-preference mapping. The six question weights are

$$\mathbf{w} = (w_{\text{peak}}, w_{\text{slope}}, w_{\text{rise}}, w_{\text{fall}}, w_{\text{boundary}}, w_{\text{context}}).$$

They are mapped to five effective region weights. The generic slope weight represents a preference for frames where query relevance changes quickly; in the five-region partition, this evidence is represented by BOUNDARY, so the slope weight is added to the boundary weight:

$$\begin{aligned} W_{\text{peak}} &= w_{\text{peak}}, \\ W_{\text{rising}} &= w_{\text{rise}}, \\ W_{\text{falling}} &= w_{\text{fall}}, \\ W_{\text{boundary}} &= w_{\text{boundary}} + w_{\text{slope}}, \\ W_{\text{background}} &= w_{\text{context}}. \end{aligned}$$

Each region receives

$$n_r = \text{round}\left(N \frac{W_r}{\sum_{r'} W_{r'}}\right),$$

capped by its available number of frames. Any remaining budget is redistributed to available regions in descending effective-weight order.

Budget-aware sub-bin selection. For RISING and FALLING, sub-bin selection is used only when both the allocated budget and the region length are at least 3. The region indices are sorted by time and split into n_r approximately equal bins, where n_r is the region budget. From each bin, RIDGE selects the frame with the largest $\bar{s}_t$. If $n_r < 3$ or the region contains fewer than 3 frames, the method falls back to ranked selection by $\bar{s}_t$.

Default hyperparameters. Unless otherwise stated, we use $\sigma = 2.0$, $p_{\min} = 0.15$, $w_0 = 3$, $w_{\max} = 10$, $k_s = 0.5$, $k_e = 4.0$, $k_b = 2.0$, and a minimum temporal gap of 2 for ranked selection regions.

B Weight-generation Prompt

When RIDGE uses question-specific region weights, the auxiliary LLM receives only the video question and the instruction below. It returns a six-dimensional preference vector in JSON, which the allocation step parses deterministically.

```
You are an expert in video analysis. Given a
    video question,
analyze what visual evidence is needed to
    answer it, and then
output a weight vector that controls how
    frames are selected
from the video's query-relevance curve.

Video question:
{question}

Weight dimensions:
- peak_similarity: Are frames highly similar
    to the question
  needed? (High for descriptive questions,
    e.g.,
  "What is shown in the scene?")
- slope_abs: Are frames with rapid changes
    needed?
  (High for dynamic or procedural questions.)
- rising_slope: Are frames from an increasing
    similarity trend
  needed? (High for causal or setup questions,
    e.g., "Why...?")
- falling_slope: Are frames from a decreasing
    similarity trend
  needed? (High for follow-up or consequence
    questions, e.g.,
  "What happens afterward?")
- boundary_change: Are event-boundary or
    abrupt-change frames
  needed? (High for questions about turning
    points.)
- context_density: Are frames from
    information-dense regions
  needed? (High for questions requiring
    context.)

Return strictly in the following JSON format.
    Each weight
must be an integer from 0 to 10:
{
  "peak_similarity": <int>,
  "slope_abs": <int>,
  "rising_slope": <int>,
  "falling_slope": <int>,
  "boundary_change": <int>,
  "context_density": <int>,
  "reasoning": "<brief rationale>"
}

Output JSON only. Do not output any other
    content.
```

C Compared Frame-selection Methods

Uniform sampling. Uniform sampling is the non-query-aware reference in the main comparison. It places a fixed frame budget at evenly spaced temporal locations and therefore preserves coarse coverage without requiring a scoring model or an additional selection algorithm. Its limitation is that the same temporal allocation is used for every question: sparse query-relevant evidence may fall between samples, while visually uninformative spans still consume frames.

AKS. Adaptive Keyframe Sampling (AKS) is a plug-in keyframe selector designed to trade off prompt relevance against temporal coverage (Tang et al., 2025a). AKS obtains prompt–frame matching scores from a lightweight vision–language model and approximates coverage with a recursive temporal binning objective. Its adaptive procedure contrasts two extreme policies—top-score sampling, which can collapse into a narrow relevant interval, and binned sampling, which enforces coverage more strongly—and recursively decides whether a temporal bin should be split or should directly return its top-scoring frames. The selected set therefore remains question-aware while avoiding a purely peak-centered allocation when coverage is beneficial.

BOLT. BOLT studies training-free frame selection for off-the-shelf long-video VLMs and adopts query-guided inverse transform sampling as its strongest selection strategy (Liu et al., 2025). It first builds a frame probability distribution from query–frame similarity scores, with a sharpness parameter controlling how strongly the distribution favors high scores. Sampling from the inverse cumulative distribution gives highly relevant frames a larger selection probability while retaining nonzero probability for lower-score frames. This design is intended to reduce the redundancy of deterministic top-k selection and retain more temporal context and diversity at inference time without training the downstream VLM.

MDP^3. MDP^3 formulates frame selection as a list-wise problem that should jointly respect query relevance, diversity among selected frames, and temporal sequentiality (Sun et al., 2025). It constructs a query-conditioned similarity matrix through a conditional Gaussian kernel in an RKHS and uses a determinantal point process to favor relevant yet non-redundant frame subsets. To preserve sequential structure, the video is segmented and the DPP selection for each segment is conditioned on the previous segment. The resulting selection-size allocation is modeled as a Markov decision process

and solved with dynamic programming, yielding a training-free selector that explicitly discourages list-wise redundancy.

FOCUS. FOCUS is a training-free, model-agnostic keyframe selector that treats short temporal clips as arms in a budgeted combinatorial pure-exploration problem (Zhu et al., 2025b). Each clip maintains an empirical relevance estimate and a Bernstein-style confidence radius, so the method can spend scoring budget on clips that are either promising or uncertain. Its practical two-stage schedule first performs coarse parallel exploration over clips and then applies batched optimistic exploration to the strongest candidates before choosing clips by their empirical means. Within selected clips, FOCUS constructs frame-level selection distributions from observed and interpolated relevance rewards to produce the final keyframes under the token budget.

D Evaluation Benchmarks

Video-MME. Video-MME is a broad benchmark for evaluating multimodal LLMs on video analysis (Fu et al., 2025). It contains 900 manually curated videos and 2,700 multiple-choice question–answer pairs across six visual domains and 30 fine-grained categories. The videos span short, medium, and long regimes from 11 seconds to one hour, and the benchmark also provides subtitles and audio to probe multi-modal evidence use. This coverage makes Video-MME useful for testing frame selection across both compact clips and longer temporal contexts; in our experiments, we use the video-only setting without subtitles.

LongVideoBench. LongVideoBench evaluates long-context interleaved video–language understanding on 3,763 web videos with subtitles and 6,678 human-annotated multiple-choice questions (Wu et al., 2024). Its central task is referring reasoning: a referring query identifies one or more relevant video contexts, and the model must retrieve and reason over details in those contexts to answer the question. The benchmark separates perception questions from relation questions and further organizes them into 17 categories over videos ranging from seconds to hour-long inputs. It is therefore sensitive to whether a selector keeps the specific frames needed for detailed long-context reasoning.

MLVU. MLVU is a multi-task benchmark built to diagnose long-video understanding across video genres, durations, and task formats (Zhou et al., 2025). Its videos span roughly three minutes to more than two hours with an average duration of about 12 minutes, covering real-world sources such as movies, documentaries, TV series, egocentric videos, and surveillance footage as well as animated and game videos. MLVU defines nine task categories, including topic reasoning, anomaly recognition, summarization, needle QA, ego reasoning, plot QA, sub-scene captioning, action counting, and action ordering. These tasks mix holistic, single-detail, and multi-detail demands, making the benchmark informative for selectors that must balance global coverage with sparse local evidence.

LVBench. LVBench targets extreme long-form video understanding (Wang et al., 2025a). Its 103 publicly sourced videos are selected from long videos with coherent structure, multiple chronological events, and content that can be understood visually without relying on audio; the benchmark contains 1,549 manually annotated QA pairs with an average video duration exceeding one hour. Its questions probe six core capabilities: key information retrieval, event understanding, entity recognition, reasoning, temporal grounding, and summarization. Because the evaluated videos are long and multi-event, LVBench exposes whether a frame selector preserves evidence needed for long-horizon retrieval and temporal reasoning rather than only salient isolated moments.

E Complexity and Runtime Analysis

RIDGE introduces only post-processing on a precomputed similarity sequence and requires no iterative LVLM calls. Stage 1 (smoothing, finite differences, peak detection, and region labeling) runs in $\mathcal{O}(T)$. Stage 2 operates over a fixed set of structural regions. Stage 3 is dominated by ranked selection and backfilling, giving a worst-case complexity of $\mathcal{O}(T \log T)$. This is asymptotically comparable to a plain top-k selector and cheaper than LVLM-iterative selectors such as A.I.R., which require additional model calls.

Table 9 reports wall-clock selection time per video on LVB, excluding the shared cost of computing s_t. RIDGE adds only 8.1 ms of post-processing, remains close to lightweight non-iterative selectors, and is far faster than LVLM-iterative methods. The optional Qwen3-8B weight call is performed once per input question at inference time, adding a small

Table 9: Per-video selection latency on LongVideoBench at N=32. Time excludes the shared $\{s_t\}$ computation. § denotes LVLM-iterative methods.

Method	Selection Time	LVLM Calls
Uniform	< 1 ms	0
AKS	5.2 ms	0
BOLT	12.4 ms	0
A.I.R.§	3.4 s	≥ 1
RIDGE (w/o LLM)	8.1 ms	0
RIDGE (w/ LLM)	8.1 ms + 0.6 s*	0

* Auxiliary LLM call is performed once per input question at inference time.

question-side preprocessing cost; however, it does not introduce additional downstream LVLM calls.

F Hyperparameter Sensitivity

Beyond the smoothing scale σ analyzed in Section 4.3, the region segmentation stage of RIDGE contains four core hyperparameters: the peak prominence threshold $p_{\min}$, which determines how many local maxima are retained as event cores; the base peak half-width w_0, which specifies the anchor size of the curvature-adaptive peak window; the slope threshold coefficient k_s, which controls the temporal extent of the rising and falling regions around each event core; and the boundary threshold coefficient k_b, which affects the detection of abrupt changes among residual background frames.

Table 10 reports the sensitivity of RIDGE to these hyperparameters on LVB and MLVU. For each sweep, one hyperparameter is varied while the remaining three are fixed to their default values. Across all tested configurations, RIDGE remains substantially above the uniform sampling baseline, which achieves 58.9% on LVB and 57.3% on MLVU. This indicates that the performance gain of RIDGE is not tied to a narrow hyperparameter setting.

Among the four hyperparameters, k_s exhibits the largest variation, with ranges of 1.4 percentage points on LVB and 1.1 percentage points on MLVU. This is expected, as k_s directly determines how much temporal context is included around each detected event core. In contrast, k_b varies by 0.7 percentage points on LVB and 1.1 percentage points on MLVU. This suggests that boundary frames play a useful but secondary role, since they usually occupy only a small fraction of the sampling budget. The peak prominence threshold $p_{\min}$ and the base half-width w_0 show moderate sensitivity. Very small $p_{\min}$ values may introduce spurious event cores, whereas overly large values may suppress weaker but informative events. Similarly, too small or too large values of w_0 can under-cover or over-cover local event regions. Overall, the default setting provides the best average performance while maintaining stable behavior under reasonable perturbations.

Table 10: Hyperparameter sensitivity of RIDGE on LVB and MLVU with Qwen2.5-VL-7B (N=32). Accuracy (%) is reported. The default configuration is highlighted. Bold entries in the first four columns indicate the hyperparameter being varied.

$p_{\min}$	w_0	k_s	k_b	LVB	MLVU
0.15	3	0.5	2.0	65.4	69.6
0.05	3	0.5	2.0	65.1	69.2
0.10	3	0.5	2.0	64.8	69.3
0.30	3	0.5	2.0	64.4	69.7
0.15	**1**	0.5	2.0	64.6	69.0
0.15	**5**	0.5	2.0	65.1	69.8
0.15	**8**	0.5	2.0	64.2	68.6
0.15	3	**0.3**	2.0	64.5	69.8
0.15	3	**0.8**	2.0	64.9	69.2
0.15	3	**1.0**	2.0	64.0	68.7
0.15	3	0.5	**1.5**	64.7	70.1
0.15	3	0.5	**2.5**	65.2	69.4
0.15	3	0.5	**3.0**	64.8	69.0

We additionally evaluate two safety-clamp parameters: the maximum peak half-width $w_{\max}$=10 and the maximum extension factor k_e=4.0. These parameters are rarely activated in practice. Varying $w_{\max}$ within $[6, 15]$ and k_e within $[2.0, 6.0]$ changes accuracy by less than 0.5 percentage points on both benchmarks; therefore, we omit the detailed sweep.

G Multi-Peak Analysis

RIDGE keeps multiple candidate peaks rather than restricting selection to the global maximum. We analyze this behavior on LongVideoBench with Qwen2.5-VL-7B and N=32, comparing with uniform sampling and AKS, the strongest competing selector on this benchmark in Table 1. We partition instances according to whether RIDGE's default peak detector returns one or multiple peaks.

RIDGE improves over AKS by 2.80 percentage points on the multi-peak subset, compared with 0.39 percentage points on the single-peak subset. This larger gain is consistent with RIDGE retaining evidence beyond the dominant peak. On the

Table 11: Performance on single-peak and multi-peak LongVideoBench subsets with Qwen2.5-VL-7B and N=32. Accuracy (%) is reported.

Subset	Uniform	AKS	RIDGE	Δ vs. Uniform	Δ vs. AKS
All	58.94	63.20	65.37	+6.43	+2.17
Single-peak	71.51	75.11	75.50	+3.99	+0.39
Multi-peak	54.46	58.96	61.76	+7.30	+2.80

Table 12: Held-out hyperparameter control with Qwen2.5-VL-7B and N=32. Accuracy (%) is reported. Avg. excludes LVB-dev and averages the four evaluation sets.

Benchmark	Uniform	Default	Held-out
LVB-dev	57.7	68.2	69.4
LVB-held-out	59.3	64.7	64.2
Video-MME	61.2	63.7	64.3
MLVU	59.7	69.6	69.4
LVBench	38.5	50.9	51.4
Avg.	54.7	62.2	62.3

non-short multi-peak subset, an average of 7.15% of RIDGE-selected frames fall inside the strongest-peak window. Thus, the selections are not concentrated in that window, although RIDGE does not guarantee that every detected peak receives a selected frame.

H Held-out Hyperparameter Control

To assess whether RIDGE depends on evaluation-set tuning, we create a fixed, video-disjoint development/evaluation split of LongVideoBench. Using only the 20% development subset, we select among 16 pre-declared configurations, freeze the selected configuration, and evaluate it on the disjoint 80% LongVideoBench subset and the other three benchmarks without further adaptation. The paper-default configuration was fixed before all sensitivity sweeps and is evaluated under the same protocol.

The configuration selected on LVB-dev uses σ=2.0, $p_{\min}$=0.15, w_0=3, k_s=0.5, k_e=4.0, and k_b=2.5; the safety clamps $w_{\max}$=10 and the minimum temporal gap of 2 remain fixed. It achieves essentially the same four-benchmark average as the paper default (62.3 vs. 62.2) and consistently outperforms uniform sampling across all four evaluation sets. These results show that RIDGE's gains are preserved when its hyperparameters are selected without access to the reported evaluation labels.

I More Visualization Results

Figure 6 presents six representative examples spanning diverse question types—action recognition, text reading, object identification, color discrimination, movie reference, and counting—to illustrate the broad applicability of RIDGE. In each case the base model selects frames that miss the decisive visual evidence, leading to plausible but incorrect answers.

For instance, Qwen2.5-VL-7B misreads "No smoking" as "Safety first" and mistakes tires for boxes, while LLaVA-OV-7B confuses the parodied movie and miscounts the number of women off-stage. After integrating RIDGE, the key-frame selection shifts to the query-relevant moments (marked by green stars), and all six answers are corrected. These results suggest that RIDGE generalizes across the evaluated backbones and improves performance across diverse visual reasoning skills.

Figure 7 compares RIDGE with uniform sampling, AKS (Tang et al., 2025a), and FOCUS (Zhu et al., 2025b), providing additional qualitative evidence for the effectiveness of RIDGE. The two examples illustrate a common failure mode in long-video question answering: the visually decisive evidence is not always contained in a single high-relevance frame. Instead, answering the question often requires a short temporal process, including what happens before the event, the event core itself, and what happens afterward.

Figure 6: Qualitative examples showing how RIDGE improves long-video question answering with Qwen2.5-VL-7B and LLaVA-OV-7B. Each panel presents a question, sampled video frames, and the answers produced with and without RIDGE. Green stars mark key frames selected by RIDGE, while red crosses and green check marks denote incorrect and correct answers, respectively. Without RIDGE, the models may miss fine-grained visual evidence and produce plausible but incorrect answers. RIDGE shifts frame selection toward query-relevant moments, helping the models recover the evidence needed to answer the questions correctly.

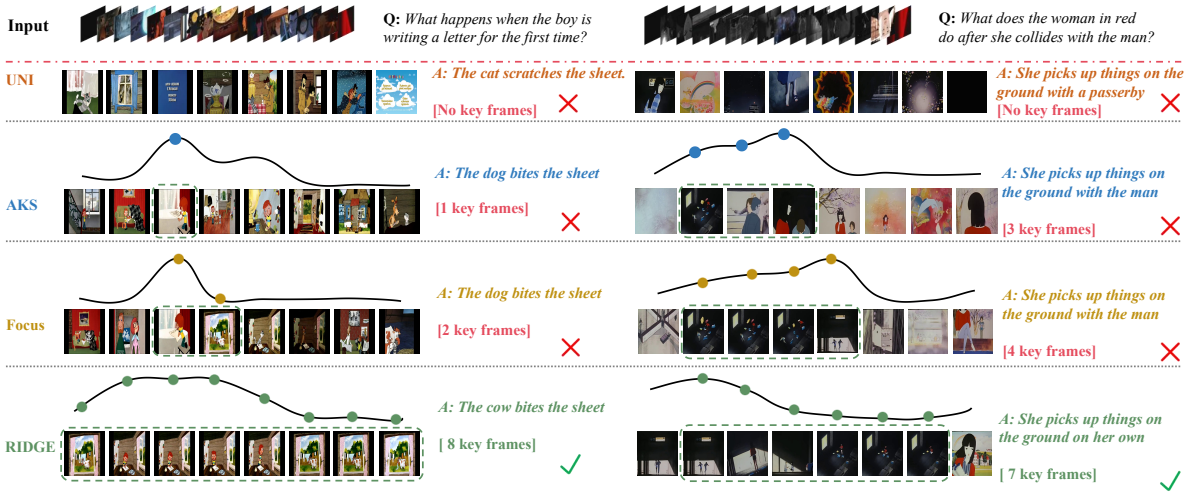

Figure 7: Qualitative comparison between uniform sampling, AKS, FOCUS, and RIDGE on long-video question answering. The top row shows the input video frames and the query. For each method, the dashed green boxes indicate the selected key frames or selected temporal spans, and the curve shows the query-frame relevance trajectory. Red crosses denote incorrect answers and green check marks denote correct answers. Compared with prior keyframe selection methods, RIDGE explicitly uses the shape of the relevance curve to preserve event cores, pre-event buildup, aftermath, and transition evidence, which leads to more faithful visual grounding.