RACER: REFLECTIVE AGENT COUPLING QUERY INTERPRETATION AND TOOL-BASED RETRIEVAL FOR FRAME SELECTION IN LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models (Vid-LLMs) excel at diverse video-language tasks by reasoning over selected frames. However, frame selection for long videos remains challenging, as it requires retrieving relevant frames distributed across segments from a large candidate pool given complex queries. This paper investigates dominant approaches to long-video frame selection from a task-decomposition perspective, identifying two key challenges: the Query Comprehension Gap in similarity-based methods and the Interpretation–Selection Gap in judgment-based methods. To address them, we propose RACER, a training-free reflective agentic framework that decomposes long-video frame selection into query interpretation driven by a lightweight Vid-LLM and evidence localization supported by an embedding model serving as a retrieval tool. Specifically, the Vid-LLM is responsible solely for reformulating the complex query into sub-queries that make implicit information requirements explicit, mitigating the Query Comprehension Gap. Meanwhile, the retrieval tool leverages these sub-queries to localize relevant evidence, relieving the Vid-LLM of direct frame selection and thus addressing the Interpretation–Selection Gap. Finally, the retrieved frames are fed back to the Vid-LLM for sub-query refinement, forming a reflection loop that iteratively improves query interpretation and frame selection. Experiments across multiple benchmarks show that RACER consistently improves long video understanding. Notably, RACER achieves effective frame selection even with limited-capability components, demonstrating that agentic integration enables these components to enhance more capable Vid-LLMs. *Code will be released.*

1 INTRODUCTION

Video large language models (Vid-LLMs) (Zhang et al., 2023; Maaz et al., 2024; Lin et al., 2024; Yan et al., 2026) integrate selected video frames with textual queries, achieving significant advances across diverse video-language tasks. However, key-frame selection in long videos remains challenging, as query-relevant evidence is scattered across segments within a vast candidate frame pool, while complex queries often fail to specify the required visual evidence. These challenges pose reliability and safety risks in autonomous driving (Ma et al., 2024), robotics (Wang et al., 2025a), and security (Lv & Sun, 2024).

Various approaches have been proposed for long-video frame selection, broadly following similarity-based and judgment-based paradigms. Similarity-based methods use embedding models (e.g., CLIP (Radford et al., 2021)) to score query-frame similarity for frame selection. BOLT (Liu et al., 2025) improves selection through inverse transform sampling, while Q-Frame (Zhang et al., 2025c) combines Gumbel-Max sampling with multi-resolution adaptation. AKS (Tang et al., 2025) incorporates temporal coverage into selection, whereas MDP^3 (Sun et al., 2025) considers diversity and sequentiality in frame-set construction. In contrast, judgment-based methods employ VLM/LLM decisions on candidate frames and segments, either assigning relevance levels or performing binary selection. Existing approaches include LLM assessment of VLM-generated captions (Wang et al., 2025b; Zhang et al., 2024a), iterative LLM selection of temporal search regions (Wang et al., 2024a), and iterative VLM assignment of relevance levels to promising candidates (Zou et al., 2026). However, existing methods treat frame selection as a single task driven by

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

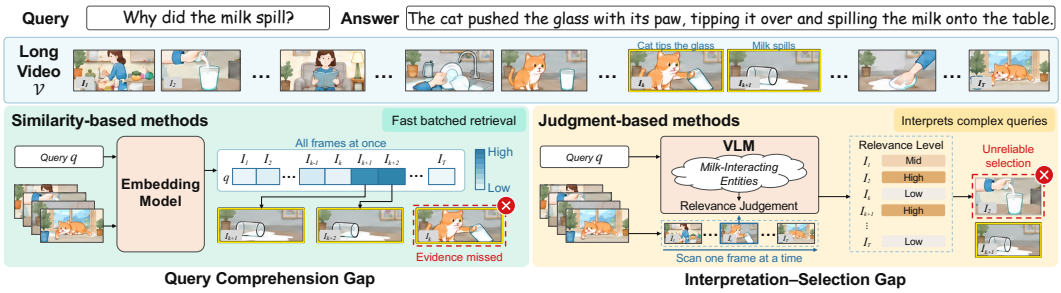


Figure 1: **Two key challenges in long-video frame selection.** Answering "Why did the milk spill?" requires evidence of the cat tipping the glass and the resulting spill. **Query Comprehension Gap (left):** similarity-based methods overlook implicit causal requirements, retrieving spilled-milk frames while missing the cat tipping the glass. **Interpretation–Selection Gap (right):** judgment-based methods understand implicit evidence requirements but fail to perform selection reliably, choosing a milk-pouring frame instead of the causal action.

similarity scoring or VLM/LLM judgment, leaving the processes of determining what evidence is needed and locating where it appears largely coupled.

This paper investigates long-video frame selection from a task-decomposition perspective, identifying the **Query Comprehension Gap** in similarity-based methods when interpreting evidence requirements and the **Interpretation–Selection Gap** in judgment-based methods when localizing relevant evidence, as illustrated in Figure 1. Similarity-based methods enable efficient and precise retrieval from large candidate pools, but embedding models may fail to capture the temporal, causal, and implicit information requirements of complex queries. Judgment-based methods can interpret complex queries and describe potentially relevant visual evidence, yet such understanding does not necessarily translate into reliable frame or segment selection.

To bridge these gaps, we propose **RACER**, a training-free reflective agentic framework that decomposes long-video frame selection into query interpretation assigned exclusively to a lightweight Vid-LLM and evidence localization delegated to an embedding model serving as a retrieval tool. Specifically, the Vid-LLM observes a video overview and reformulates the complex query into sub-queries that make distinct implicit information requirements explicit to the embedding model, thereby mitigating the Query Comprehension Gap. Meanwhile, the retrieval tool grounds the original query and the generated sub-queries in candidate frames through similarity scoring, relieving the Vid-LLM of direct frame selection and addressing the Interpretation–Selection Gap. Finally, the retrieved frames are fed back to the Vid-LLM for sub-query refinement, forming a reflection loop that iteratively improves query interpretation and frame selection.

Experiments demonstrate that RACER consistently improves performance on long-video understanding benchmarks, including Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), and LongVideoBench (Wu et al., 2024). These gains extend across multiple Vid-LLMs, including LLaVA-OneVision (Li et al., 2025a), LLaVA-Video (Zhang et al., 2024b), Qwen2.5-VL (Bai et al., 2025b), and Qwen3-VL (Bai et al., 2025a). Notably, RACER achieves effective key-frame selection even with limited-capability components, demonstrating that agentic integration enables these components to enhance more capable Vid-LLMs.

Our contributions are summarized as follows:

- We investigate long-video frame selection from a task-decomposition perspective, identifying the **Query Comprehension Gap** in similarity-based methods and the **Interpretation–Selection Gap** in judgment-based methods.
- We propose **RACER**, a training-free reflective agentic framework that decomposes long-video frame selection into query interpretation assigned exclusively to a lightweight Vid-LLM and evidence localization delegated to an embedding model serving as a retrieval tool.
- Comprehensive experiments demonstrate RACER's effectiveness across multiple long-video understanding benchmarks and Vid-LLM families. The results further show that agentic integration enables limited-capability components to enhance more capable Vid-LLMs.

2 Related Work

2.1 Video Large Language Models (Vid-LLMs)

Early Vid-LLMs connect video features to LLMs through query-based adapters (Li et al., 2025b; Zhang et al., 2023) or spatiotemporal pooling (Maaz et al., 2024). Video-LLaVA (Lin et al., 2024) aligns image and video representations, while VILA (Wang et al., 2024b) adopts interleaved pre-training. More recent models (Li et al., 2025a; Zhang et al., 2024b; Bai et al., 2025b;a; Zhu et al., 2025), further improve unified image-video modeling, instruction tuning, and spatiotemporal representation. For long videos, MovieChat (Song et al., 2024) and LLaMA-VID (Li et al., 2024) compress visual inputs through memory or compact tokens, while LongVA (Zhang et al., 2025b) and LongVU (Shen et al., 2025) extend context capacity through adaptive spatiotemporal compression. Nevertheless, Vid-LLMs process only limited sampled frames, making effective key-frame selection critical for video understanding.

2.2 Key-Frame Selection for Vid-LLMs

Existing Vid-LLM frame-selection methods broadly follow similarity-based and judgment-based paradigms. Similarity-based methods use embedding models, such as CLIP (Radford et al., 2021), to score query–frame similarity for frame selection. Building on these similarity scores, BOLT (Liu et al., 2025) uses inverse transform sampling, while Q-Frame (Zhang et al., 2025c) combines Gumbel-Max sampling with multi-resolution adaptation. AKS (Tang et al., 2025) adds temporal coverage, whereas MDP^3 (Sun et al., 2025) considers diversity and sequentiality. Other methods adapt frame allocation using temporal structure or similarity distributions (Miraj et al., 2026; Zhang et al., 2026). In contrast, judgment-based methods use VLM/LLM decisions to assign relevance levels or select candidate frames and temporal regions. Caption-mediated judgments support hierarchical frame selection (Wang et al., 2025b) and supporting-frame identification in grounded VideoQA (Zhang et al., 2024a). VideoAgent (Wang et al., 2024a) iteratively selects temporal search regions for progressive video analysis, while A.I.R. (Zou et al., 2026) assigns relevance levels to promising candidates and expands the search around validated frames. Other approaches learn selection decisions through localization (Yu et al., 2023), rankings derived from downstream prediction losses (Yu et al., 2025), or autoregressive policies trained with answer feedback (Lee et al., 2026). While effective, these methods primarily treat frame selection as a unified task driven by similarity scoring or VLM/LLM judgment, leaving evidence requirement interpretation and localization largely coupled, a limitation this work aims to address through task decomposition.

3 Method

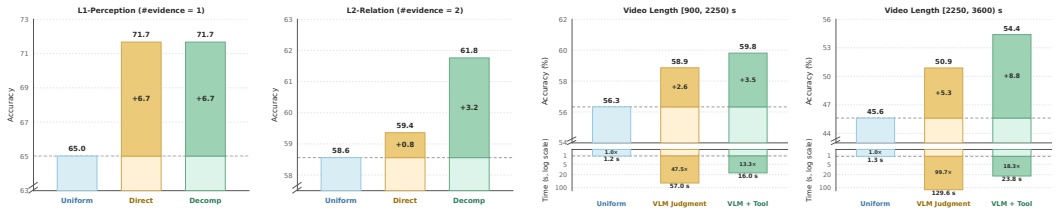
3.1 Challenges in Dominant Frame Selection Approaches

Effective frame selection requires interpreting query-implied evidence requirements and reliably localizing corresponding frames within a large candidate pool. However, similarity-based methods face the Query Comprehension Gap in interpreting complex queries, whereas judgment-based methods face the Interpretation–Selection Gap in selecting relevant frames. This section analyzes these two challenges.

3.1.1 Query Comprehension Gap in Similarity-Based Methods

Similarity-based methods select frames by query–frame embedding similarity. While effective for simple queries requiring a single salient evidence target, they are less effective for complex queries requiring multiple evidence targets linked by temporal, causal, or relational dependencies.

To reveal this challenge, two LongVideoBench (Wu et al., 2024) subsets differing in query complexity are considered: single-evidence L1-Perception and multi-evidence L2-Relation. Three methods are compared: uniform sampling, direct selection using only the original query, and decomposed selection using both the original query and sub-queries generated through query decomposition by a lightweight Vid-LLM (Qwen2.5-VL-3B) observing 32 uniformly sampled overview frames. Both selection methods select the top-32 frames by CLIP query–frame similarity. All methods use 32 frames for VideoQA with LLaVA-OneVision. As shown in Figure 2a, direct and decomposed



(a) **Query Comprehension Gap.** Accuracy of Uniform Sampling, Direct (CLIP selection with the original query), and Decomp (CLIP selection with the original query and VLM-generated sub-queries) on single-evidence (left) and multi-evidence (right) queries.

(b) **Interpretation–Selection Gap.** Accuracy (top) and time cost (bottom) of Uniform Sampling, VLM Judgment (VLM-assigned frame relevance levels), and VLM+Tool (VLM query decomposition with CLIP frame selection) across two video-duration groups.

Figure 2: **Empirical analysis of two gaps on LongVideoBench subsets. (a)** Embedding models such as CLIP handle single-evidence queries effectively but struggle with implicit requirements in complex queries, which query decomposition helps make explicit. **(b)** VLMs are more effective at interpreting complex queries and describing relevant evidence than at directly selecting frames, as using these interpretations to guide tool-based selection yields higher accuracy and lower latency than direct VLM selection, especially for long videos.

selection yield equally substantial accuracy gains over uniform sampling on single-evidence L1-Perception queries, whereas on multi-evidence L2-Relation queries, the gain is only 0.8 for direct selection compared with 3.2 for decomposed selection. These results illustrate the **Query Comprehension Gap** in similarity-based methods: embedding models struggle to capture implicit requirements in complex queries, while query decomposition makes these requirements explicit.

3.1.2 Interpretation–Selection Gap in Judgment-Based Methods

Judgment-based methods employ VLM/LLM decisions on candidate frames and segments. While effective at understanding complex queries and describing potentially relevant evidence, these models can be inefficient and unreliable when assigning relevance levels or making binary selections over large candidate pools.

To reveal this challenge, the subset of LongVideoBench (Wu et al., 2024) with the widest duration range is considered and divided by video duration into two groups: 900–2250 and 2250–3600 seconds. Three methods are compared: uniform sampling, VLM judgment-based selection assigning a relevance level to each frame, and tool-augmented VLM selection, where the VLM only decomposes the query into sub-queries based on 32 uniformly sampled overview frames, while CLIP serves as a retrieval tool to score frames against the original query and sub-queries. Both VLM variants use Qwen2.5-VL-3B and select the top-32 frames by relevance levels or similarity scores. All methods use 32 frames for VideoQA with LLaVA-OneVision. As shown in Figure 2b, both VLM variants improve accuracy over uniform sampling, while tool-augmented VLM selection consistently outperforms VLM judgment-based selection in accuracy and efficiency, achieving an additional accuracy gain of 3.5 and a $5.4\times$ speedup in the longer-duration group. These results illustrate the **Interpretation–Selection Gap** in judgment-based methods: VLM/LLM judgment does not ensure reliable and efficient frame selection. The higher accuracy and efficiency of retrieval guided by VLM-generated interpretations suggest that VLMs are better suited to interpreting evidence requirements than selecting frames, especially for longer videos.

3.2 RACER: Reflective Agent Coupling Query Interpretation and Tool-Based Retrieval

Building on these observations, we propose **RACER**, a training-free reflective agentic framework that decomposes long-video frame selection into query interpretation and evidence localization, addressing the Query Comprehension Gap and the Interpretation–Selection Gap, respectively. As illustrated in Figure 3, RACER begins with query interpretation: a lightweight Vid-LLM observes a video overview and decomposes the query into sub-queries, making distinct implicit information requirements explicit to the embedding model. For evidence localization, the embedding model serves

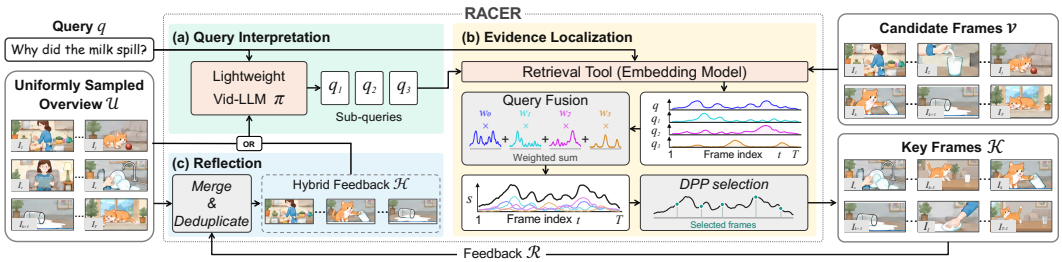


Figure 3: **Overview of RACER.** RACER begins with **(a) Query Interpretation**, where a lightweight Vid-LLM uses a uniformly sampled overview to decompose the query into complementary sub-queries, making implicit evidence requirements explicit. Then, **(b) Evidence Localization** employs an embedding-based retrieval tool to score candidate frames against the original query and sub-queries, followed by entropy-adaptive score fusion and DPP selection to obtain relevant and visually diverse key frames. Finally, **(c) Reflection** merges feedback frames with the overview, removing duplicates and the least visually aligned overview frames to maintain the frame budget. The resulting hybrid overview guides sub-query refinement for the next selection round.

as a retrieval tool, grounding the original query and sub-queries in candidate frames through similarity scoring and relieving the Vid-LLM of direct frame selection. Finally, retrieved frames are fed back to the Vid-LLM for sub-query refinement, forming a reflection loop that iteratively improves query interpretation and frame selection.

3.2.1 Task Formulation: Frame Selection in Long Video Understanding

Frame selection enables Vid-LLMs with limited frame budgets to understand long videos. Given candidate frames $\mathcal{V} = \{I_t\}_{t=1}^{T}$ and a budget $N \ll T$, a selector $\mathcal{S}$ selects key frames $\mathcal{K}$ based on a query q:

$$\mathcal{K} = \mathcal{S}(\mathcal{V}, q), \qquad \mathcal{K} \subset \mathcal{V},\ |\mathcal{K}| = N. \tag{1}$$

The selected frames $\mathcal{K}$ and query q are then provided to a Vid-LLM $\mathcal{F}$ to generate an answer:

$$A = \mathcal{F}(\mathcal{K}, q). \tag{2}$$

3.2.2 Query Interpretation

As discussed in Section 3.1.1, the Query Comprehension Gap limits embedding models in handling complex queries requiring interrelated evidence. To address this gap, the Query Interpretation module decomposes the original query into sub-queries, making implicit retrieval needs explicit.

Specifically, the module employs a lightweight Vid-LLM π to generate sub-queries $\{q_1, \ldots, q_M\}$ from query q and a video overview comprising N frames uniformly sampled from the candidate pool $\mathcal{V}$:

$$\{q_1, \ldots, q_M\} = \pi(q, \mathrm{Uniform}(\mathcal{V}, N))\,. \tag{3}$$

The decomposition is guided by the prompt in Table 1, which instructs π to generate complementary, visually grounded sub-queries, each targeting a distinct information requirement implicit in the original query.

3.2.3 Tool-Based Evidence Localization

As discussed in Section 3.1.2, the Interpretation–Selection Gap limits the reliability of direct VLM relevance judgments, while inference costs increase with the candidate pool size. To address this gap, the Tool-Based Evidence Localization module delegates frame selection to an embedding model, using query–frame similarity to retrieve relevant frames from the candidate pool.

Specifically, the embedding model separately matches the original query q and generated sub-queries $q_1, \ldots, q_M$ against each candidate frame I_t through query–frame similarity:

$$S(q, I_t) = \cos(E_t(q), E_v(I_t))\,, \qquad S(q_m, I_t) = \cos(E_t(q_m), E_v(I_t))\,, \quad m = 1, \ldots, M, \tag{4}$$

where E_t and E_v denote the text and visual encoders of the embedding model, respectively.

Table 1: Prompt for Query Interpretation.

Query Interpretation Prompt

You are analyzing a video question. To answer the main question well, you sometimes need to first know the answers to simpler sub-questions about the video. Watch the video and propose 2–3 such sub-questions.

Each sub-question must:

- Be answerable by looking at a single still frame—ask about visible objects, people, places, attributes, settings, or counts. Avoid asking about actions, sequences, order, changes over time, or "what happens next"—those cannot be answered from one frame.
- Point to information the main question implicitly depends on, but does not itself ask about directly.
- Be phrased in plain natural language as a question rather than a statement or caption.
- Be different from the main question and from each other.

Main question: {query}

Table 2: Prompt for Sub-Query Refinement.

Sub-Query Refinement Prompt

You are refining sub-questions for a video question. Given the current sub-questions and newly retrieved video frames, revise the sub-questions to better capture the information needed to answer the main question.

Each refined sub-question must:

- Use the retrieved visual evidence to preserve useful sub-questions, revise ineffective ones, or introduce missing information needs.
- Be answerable by looking at a single still frame—ask about visible objects, people, places, attributes, settings, or counts. Avoid asking about actions, sequences, order, changes over time, or "what happens next".
- Point to information the main question implicitly depends on, but does not itself ask about directly.
- Be different from the main question and from each other.

Main question: {query}
Current sub-questions: {sub_queries}

The resulting similarity scores are adaptively fused into a relevance score s_t for each candidate frame:

$$s_t = w_0 S(q, I_t) + \sum_{m=1}^{M} w_m S(q_m, I_t), \qquad w_m = (1 - w_0) \frac{\exp(-\beta H_m)}{\sum_{j=1}^{M} \exp(-\beta H_j)}. \tag{5}$$

Here, w_0 is the fixed weight for the original query, w_m is the adaptive weight for sub-query q_m, and H_m is the entropy of its query–frame similarity distribution. The parameter β controls entropy sensitivity, with larger weights assigned to lower-entropy, more discriminative retrieval signals.

Finally, DPP-based sampling (Macchi, 1975; Chen et al., 2018) selects N relevant and visually diverse frames by greedily approximating determinant maximization over $L_{\mathcal{K}}$, the submatrix of the DPP kernel L indexed by $\mathcal{K}$:

$$\mathcal{K} \approx \arg\max_{|\mathcal{K}|=N} \det(L_{\mathcal{K}}), \qquad L_{ij} = (s_i s_j)^{\alpha} \frac{1 + \cos(E_v(I_i), E_v(I_j))}{2}. \tag{6}$$

Here, the kernel L combines min-max-normalized relevance scores s_t with rescaled cosine similarity to capture visual redundancy. Larger α emphasizes relevance, while smaller α favors visual diversity.

3.2.4 Reflection Loop

To iteratively enhance Query Interpretation and Tool-Based Evidence Localization, RACER establishes a reflection loop in which tool-selected key frames serve as visual feedback for sub-query refinement.

Specifically, in each reflection round, the first B frames selected by DPP-based sampling form the feedback set $\mathcal{R}$. A hybrid overview $\mathcal{H}$ combines all feedback frames with selected uniform overview frames under the budget N:

$$\mathcal{H} = \mathcal{R} \cup \mathrm{Top}_{N-|\mathcal{R}|} \left(\mathrm{Uniform}(\mathcal{V}, N) \setminus \mathcal{R};\, r_u \right), \qquad r_u = \max_{I_r \in \mathcal{R}} \frac{1 + \cos(E_v(I_u), E_v(I_r))}{2}. \tag{7}$$

After excluding frames already in $\mathcal{R}$, each remaining overview frame I_u is scored by its maximum visual similarity to any feedback frame I_r. The operator $\mathrm{Top}_k(\cdot; r_u)$ retains the k highest-scoring frames, removing the least similar overview frames while preserving all feedback frames. Thus, $\mathcal{H}$ combines retrieved evidence with visually aligned overview context.

Finally, given the original query q, hybrid overview $\mathcal{H}$, and current sub-queries$\{q_1, \ldots, q_M\}$, the lightweight Vid-LLM π refines the retrieval targets:

$$\{q'_1, \ldots, q'_M\} = \pi(q, \mathcal{H}, \{q_1, \ldots, q_M\}). \tag{8}$$

The refinement is guided by the prompt in Table 2, which instructs π to revise sub-queries using retrieved evidence while preserving complementarity and visual grounding. The refined sub-queries guide the next Tool-Based Evidence Localization round, closing the reflection loop.

Table 3: Comparison with state-of-the-art frame-selection methods across three benchmarks and four Vid-LLMs: accuracy (%) for Video-MME and LongVideoBench, and M-Avg for MLVU. Parentheses show gains over uniform sampling (positive in green). Average is the unweighted mean; **bold**/underline mark best/second-best results per Vid-LLM. RACER achieves the best or tied-best performance across all benchmarks and answering models.

Vid-LLMs	Frame Selection Method	Video-MME (1–60 min)	MLVU (3–120 min)	LongVideoBench (1–60 min)	Average
LLaVA-OV-7B	Uniform	58.5 (+0.0)	62.4 (+0.0)	56.6 (+0.0)	59.2 (+0.0)
	AKS (Tang et al., 2025)	58.4 (-0.1)	66.8 (+4.4)	59.3 (+2.7)	61.5 (+2.3)
	BOLT (Liu et al., 2025)	59.9 (+1.4)	66.8 (+4.4)	59.6 (+3.0)	62.1 (+2.9)
	MDP^3 (Sun et al., 2025)	60.5 (+2.0)	68.3 (+5.9)	59.0 (+2.4)	62.6 (+3.4)
	AdaQ (Zhang et al., 2026)	59.5 (+1.0)	69.4 (+7.0)	60.0 (+3.4)	63.0 (+3.8)
	GIFT (Ma et al., 2026)	61.2 (+2.7)	69.4 (+7.0)	59.6 (+3.0)	63.4 (+4.2)
	A.I.R. (Zou et al., 2026)	61.4 (+2.9)	69.3 (+6.9)	60.7 (+4.1)	63.8 (+4.6)
	RACER (ours)	**61.9** (+3.4)	**70.0** (+7.6)	**60.9** (+4.3)	**64.3** (+5.1)
LLaVA-Video-7B	Uniform	62.6 (+0.0)	64.7 (+0.0)	58.0 (+0.0)	61.8 (+0.0)
	AKS (Tang et al., 2025)	64.5 (+1.9)	65.4 (+0.7)	59.8 (+1.8)	63.2 (+1.4)
	BOLT (Liu et al., 2025)	64.1 (+1.5)	66.5 (+1.8)	59.8 (+1.8)	63.5 (+1.7)
	GIFT (Ma et al., 2026)	**65.0** (+2.4)	67.4 (+2.7)	60.2 (+2.2)	64.2 (+2.4)
	RACER (ours)	**65.0** (+2.4)	**72.0** (+7.3)	**61.3** (+3.3)	**66.1** (+4.3)
Qwen2.5-VL-7B	Uniform	60.8 (+0.0)	59.3 (+0.0)	58.1 (+0.0)	59.4 (+0.0)
	AKS (Tang et al., 2025)	62.8 (+2.0)	69.1 (+9.8)	58.8 (+0.7)	63.6 (+4.2)
	BOLT (Liu et al., 2025)	64.1 (+3.3)	66.3 (+7.0)	58.6 (+0.5)	63.0 (+3.6)
	MDP^3 (Sun et al., 2025)	63.8 (+3.0)	66.2 (+6.9)	60.0 (+1.9)	63.3 (+3.9)
	GIFT (Ma et al., 2026)	64.4 (+3.6)	68.2 (+8.9)	61.3 (+3.2)	64.6 (+5.2)
	A.I.R. (Zou et al., 2026)	**65.0** (+4.2)	67.5 (+8.2)	61.4 (+3.3)	64.6 (+5.2)
	RACER (ours)	**65.0** (+4.2)	**70.9** (+11.6)	**62.2** (+4.1)	**66.0** (+6.6)
Qwen3-VL-8B	Uniform	64.4 (+0.0)	64.4 (+0.0)	60.4 (+0.0)	63.1 (+0.0)
	BOLT (Liu et al., 2025)	65.7 (+1.3)	70.3 (+5.9)	62.2 (+1.8)	66.1 (+3.0)
	AdaQ (Zhang et al., 2026)	67.7 (+3.3)	72.0 (+7.6)	63.8 (+3.4)	67.8 (+4.7)
	RACER (ours)	**67.9** (+3.5)	**73.5** (+9.1)	**64.0** (+3.6)	**68.5** (+5.4)

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

To evaluate RACER for long video frame selection, we conduct experiments on Video-MME (w/o subtitles) (Fu et al., 2025), MLVU (Zhou et al., 2025), and LongVideoBench (Wu et al., 2024) using lmms-eval (Zhang et al., 2025a). Following official evaluation protocols, we report accuracy for Video-MME and LongVideoBench and M-Avg for MLVU. Four representative Vid-LLMs serve as answering models: LLaVA-OneVision (Li et al., 2025a), LLaVA-Video (Zhang et al., 2024b), Qwen2.5-VL (Bai et al., 2025b), and Qwen3-VL (Bai et al., 2025a), each with a frame budget of $N = 32$. RACER uses Qwen2.5-VL-3B for query interpretation and CLIP ViT-L/14 as the retrieval tool. Candidate frames are sampled at 2 fps for videos shorter than 120 seconds and 1 fps otherwise. Unless otherwise specified, we use $M = 3$ sub-queries, a feedback budget of $B = 24$, two reflection rounds, $w_0 = 0.6$, $\alpha = 0.75$, and $\beta = 2.2$. All components remain frozen, with no training required. Experiments are conducted on $8\times$A100-80GB GPUs.

4.2 COMPARISON TO THE STATE-OF-THE-ART

As shown in Table 3, RACER consistently improves long-video understanding across all four answering Vid-LLMs, achieving the best average performance in every setting. RACER yields average gains of 4.3–6.6 over uniform sampling and consistently outperforms prior frame-selection methods across the three benchmarks. The gains are particularly pronounced on MLVU, which contains the longest videos of the three benchmarks, up to 120 minutes. RACER improves over uniform sampling by 7.3–11.6 on MLVU. These results support the effectiveness of decomposing frame selection into query interpretation and evidence localization, with each task assigned to a distinct paradigm.

4.3 ABLATION STUDIES

Module Ablation. To assess each component's contribution, we conduct cumulative ablations on Video-MME with LLaVA-OV-7B as the answering model. As shown in Table 4, Tool-Based Retrieval improves overall accuracy from 58.5 to 60.6. Adding Query Interpretation raises accuracy to 61.5, and Reflection further improves it to 61.9, with the full framework achieving the best performance across all three video-duration groups. Notably, despite lower direct answering accuracy than

Table 4: Module ablation of RACER on Video-MME with components added cumulatively. Scores report accuracy (%) by duration and overall; **bold** marks column bests. Each component improves performance, with full RACER performing best. Despite weaker direct answering capability, Qwen2.5-VL-3B improves LLaVA-OV-7B when assigned Query Interpretation.

Answering Vid-LLM	Frame Selection	Retrieval Tool	Query Interpretation Model	Video-MME			
				Short	Medium	Long	Overall
Qwen2.5-VL-3B	Uniform Sampling	N/A	N/A	69.5	57.0	47.1	57.8
LLaVA-OV-7B	Uniform Sampling	N/A	N/A	70.2	57.1	48.2	58.5
	+ Tool-based Retrieval	CLIP ViT-L/14	N/A	71.2	58.7	51.8	60.6
	+ Query Interpretation	CLIP ViT-L/14	Qwen2.5-VL-3B	72.3	60.1	52.1	61.5
	+ Reflection (RACER)	CLIP ViT-L/14	Qwen2.5-VL-3B	**72.6**	**60.6**	**52.7**	**61.9**

Table 5: Component replacement on Video-MME with LLaVA-OV-7B. Acc.: accuracy (%); **bold**: best per group. RACER performs consistently across retrieval tools and Query Interpretation models.

Retrieval Tool	Query Interpretation Model	Acc. (%)
CLIP ViT-L/14	Qwen2.5-VL-3B	61.93
CLIP ViT-L/14	Qwen3-VL-4B	**62.04**
SigLIP-L/16	Qwen2.5-VL-3B	**62.26**
EVA02-CLIP-L/14	Qwen2.5-VL-3B	61.81

Table 6: Cumulative efficiency ablation on Video-MME with LLaVA-OV-7B. Acc.: accuracy (%); # Frames: mean distinct Vid-LLM input frames; Time: seconds. RACER improves accuracy with limited frame analysis.

Frame Selection	Acc. (%)	# Frames	Time (s)
Uniform Sampling	58.5	0.0	1.1
+ Tool-based Retrieval	60.6	0.0	15.1
+ Query Interpretation	61.5	32.0	22.4
+ Reflection (full RACER)	**61.9**	62.9	35.4

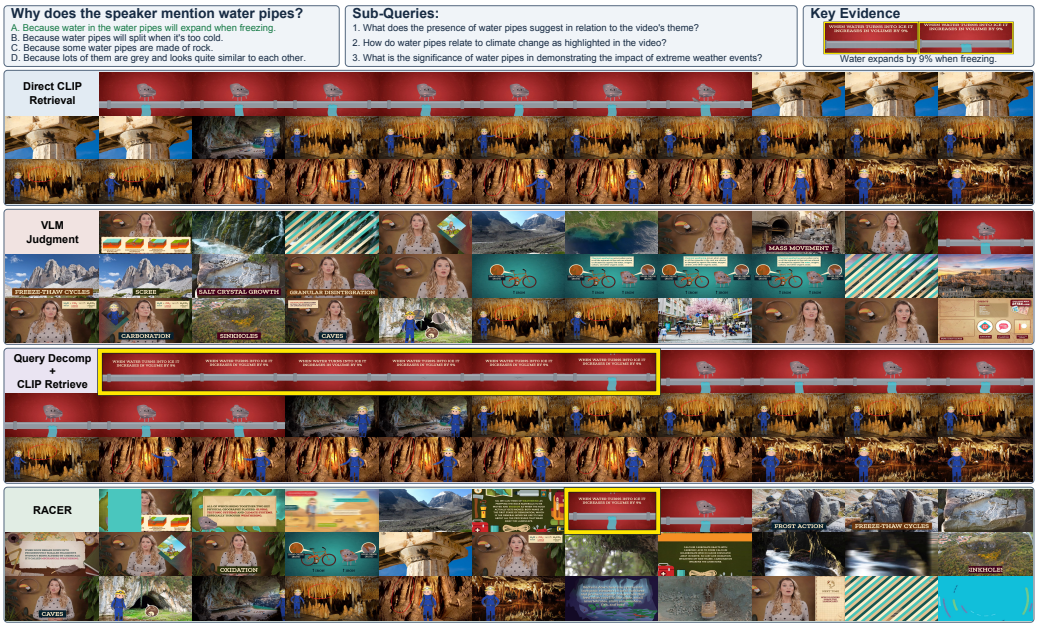
LLaVA-OV-7B (57.8 vs. 58.5 under uniform sampling), Qwen2.5-VL-3B enhances the stronger model when assigned query interpretation. These results support decomposing frame selection into query interpretation and evidence localization, with task assignments aligned with model capabilities, enabling limited-capability components to enhance more capable Vid-LLMs.

Component Ablation. To assess sensitivity to component choices, we replace the retrieval tool or Query Interpretation model while keeping the remaining framework unchanged. As shown in Table 5, replacing CLIP ViT-L/14 with SigLIP-L/16 or EVA02-CLIP-L/14 yields accuracies of 62.26 and 61.81, compared with 61.93 for the default configuration. Replacing Qwen2.5-VL-3B with Qwen3-VL-4B improves accuracy to 62.04. Comparable performance across configurations supports the applicability of RACER's task decomposition to different models for query interpretation and evidence localization.

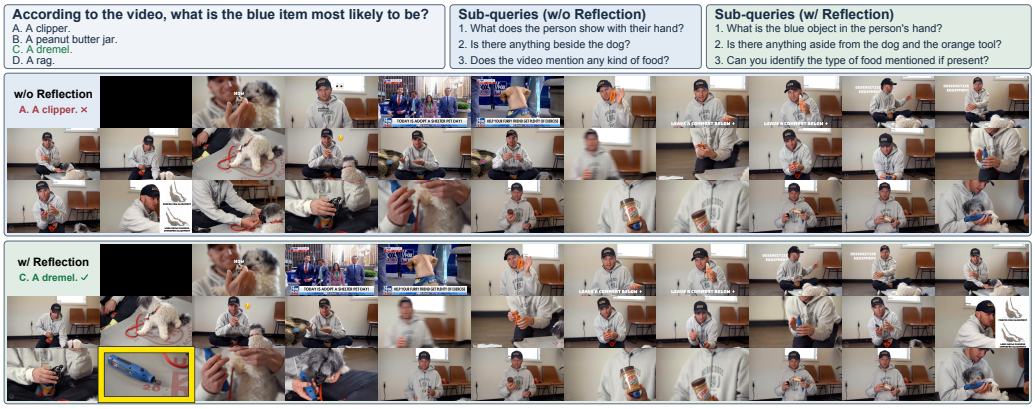
Efficiency Ablation. To analyze RACER's computational overhead, we measure accuracy, inference time, and the number of distinct frames analyzed by the lightweight Vid-LLM across all rounds as components are added cumulatively on Video-MME. As shown in Table 6, Tool-Based Retrieval takes 15.1 seconds without requiring Vid-LLM frame analysis. Query Decomposition processes 32 overview frames, increasing runtime to 22.4 seconds. With Reflection, the total reaches an average of 62.9 distinct frames and 35.4 seconds. These results quantify the additional runtime of query interpretation and reflection, while showing that candidate-pool scoring is delegated to CLIP and Vid-LLM input remains limited to a compact set of distinct frames.

4.4 Qualitative Results

Figure 4a illustrates the benefits of decomposing frame selection into query interpretation and evidence localization. Direct CLIP retrieval selects repetitive frames, whereas VLM judgment captures diverse content relevant to implicit query requirements, such as geological processes. Both, however, miss the decisive evidence. Query Interpretation enables CLIP to retrieve the explanation that water expands upon freezing, and full RACER retains this evidence with less redundancy. Beyond this division of responsibilities, reflection further improves frame selection, as shown in Figure 4b. Broad sub-queries initially miss the decisive close-up, leading to the incorrect answer, *a clipper*. Reflection refines the target to the blue object in the person's hand, distinguishing it from the dog and orange tool. The updated sub-queries recover a close-up of the tool and its sanding attachment, supporting the correct answer, *a dremel*.



(a) Query Interpretation and Tool-Based Evidence Localization.



(b) Reflection Loop.

Figure 4: **Qualitative visualization of RACER. (a)** Query interpretation guides tool-based evidence localization to recover the explanation that water expands upon freezing, missed by direct CLIP retrieval and VLM judgment. Full RACER retains this evidence with less redundancy. **(b)** Reflection refines sub-queries to target the blue object, recovering a previously missed close-up and correcting the answer from *a clipper* to *a dremel*.

5 Conclusion

This paper investigates long-video frame selection from a task-decomposition perspective, identifying the Query Comprehension Gap in similarity-based methods and the Interpretation–Selection Gap in judgment-based methods. To address these gaps, RACER decomposes frame selection into query interpretation and evidence localization within a training-free reflective agentic framework. A lightweight Vid-LLM makes implicit evidence requirements explicit through query decomposition, while an embedding-based retrieval tool selects relevant frames. Reflection on retrieved evidence refines the sub-queries, iteratively improving frame selection. Experiments across three benchmarks and four answering Vid-LLMs demonstrate consistent gains in long-video understanding. These findings support assigning query interpretation and evidence localization to models with suitable capabilities, enabling limited-capability components to enhance more capable Vid-LLMs.

Future work will explore open-ended video understanding and more efficient reflection strategies to reduce computational overhead.

AI USE STATEMENT

Generative AI assisted with two aspects of this work. First, AI assistants edited author-written drafts for grammar, clarity, and readability. Second, AI coding assistants supported method implementation and debugging based on author-defined ideas and designs. The authors are responsible for reviewing AI-assisted edits, testing code, and verifying outputs, and take full responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

RACER is a training-free framework that decomposes long-video frame selection into query interpretation and evidence localization. Experiments use publicly available benchmarks and pretrained models without additional training. RACER may inherit model biases and inaccuracies, and omitted evidence may produce incomplete or incorrect answers. Responsible use should respect privacy and dataset and model usage conditions. Human oversight remains important in safety-sensitive applications, as benchmark improvements do not guarantee reliable real-world performance.

REPRODUCIBILITY STATEMENT

RACER uses publicly available datasets and pretrained models and requires no additional training. The paper specifies model configurations, evaluation protocols, frame budgets, hyperparameters, and the iterative frame-selection procedure, including prompts for query interpretation and subquery refinement. Ablation studies examine module contributions, component choices, and computational overhead. We will release the code, experimental configurations, and per-question outputs to support reproduction and further research.

REFERENCES

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025b.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *NeurIPS*, pp. 5627–5638, 2018.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025.

Hosu Lee, Junho Kim, Hyunjun Kim, and Yong Man Ro. Refocus: Reinforcement-guided frame optimization for contextual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8291–8302, June 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025a.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: chat-centric video understanding. *Sci. China Inf. Sci.*, 68(10), 2025b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV (46)*, volume 15104 of *Lecture Notes in Computer Science*, pp. 323–340. Springer, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pp. 5971–5984. Association for Computational Linguistics, 2024.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: boost large vision-language model without training for long-form video understanding. In *CVPR*, pp. 3318–3327. Computer Vision Foundation / IEEE, 2025.

Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024.

Junpeng Ma, Sashuai Zhou, Guanghao Li, Xin Gao, Yue Cao, Hengyu Zeng, Yuxiang Yan, Zhibin Wang, Jun Song, Bo Zheng, et al. Gift: Global irreplaceability frame targeting for efficient video understanding. *arXiv preprint arXiv:2603.25072*, 2026.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *ECCV (45)*, volume 15103 of *Lecture Notes in Computer Science*, pp. 403–420. Springer, 2024.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL (1)*, pp. 12585–12602. Association for Computational Linguistics, 2024.

Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

Mehrajul Abadin Miraj, Abdul Mohaimen Al Radi, Shariful Islam Rayhan, Md Tanvir Alam, Ismat Rahman, Yu Tian, and Md Mosaddek Khan. Crest: Curvature-regulated event-centric sampling for efficient long-video understanding. *arXiv preprint arXiv:2605.09223*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, pp. 18221–18232. IEEE, 2024.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp^3: a training-free approach for list-wise frame selection in video-llms. In *ICCV*, pp. 24090–24101. IEEE, 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *CVPR*, pp. 29118–29128. Computer Vision Foundation / IEEE, 2025.

Beichen Wang, Juexiao Zhang, Shuwen Dong, Irving Fang, and Chen Feng. VLM see, robot do: Human demo video to robot action plan via vision language model. In *IROS*, pp. 17215–17222. IEEE, 2025a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *ECCV (80)*, volume 15138 of *Lecture Notes in Computer Science*, pp. 58–76. Springer, 2024a.

Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming C. Lin, and Shan Yang. Vila: Efficient video-language alignment for video question answering. In *ECCV (62)*, volume 15120 of *Lecture Notes in Computer Science*, pp. 186–204. Springer, 2024b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for LLM reasoning on long videos. In *CVPR*, pp. 3272–3283. Computer Vision Foundation / IEEE, 2025b.

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

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, et al. Internvideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*, 2026.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning to query frames for video large language models. In *ICLR*. OpenReview.net, 2025.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *EMNLP*, pp. 21715–21737. Association for Computational Linguistics, 2024a.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP (Demos)*, pp. 543–553. Association for Computational Linguistics, 2023.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. In *NAACL (Findings)*, volume NAACL 2025 of *Findings of ACL*, pp. 881–916. Association for Computational Linguistics, 2025a.

Kun Zhang, Chenxin Fang, Tao Chen, Baiyang Song, Yunhang Shen, Yiyi Zhou, and Rongrong Ji. Towards fast and effective long video understanding of multimodal large language models via adaptive quasi-gaussian sampling. *arXiv preprint arXiv:2606.24187*, 2026.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Trans. Mach. Learn. Res.*, 2025, 2025b.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *ICCV*, pp. 22056–22065. IEEE, 2025c.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: benchmarking multi-task long video understanding. In *CVPR*, pp. 13691–13701. Computer Vision Foundation / IEEE, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. Air: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, volume 2026, pp. 11302–11329, 2026.

Table 7: Accuracy (%) across six Video-MME question categories with LLaVA-OV-7B.

Method	Perception	Recognition	Reasoning	Information Synopsis	OCR Problems	Counting Problem
Uniform	68.9	60.8	53.2	74.3	66.2	36.6
CLIP Retrieve	69.2	59.5	53.9	66.9	66.2	37.3
VLM Judgment	70.7	62.1	55.8	75.2	66.2	36.2
RACER (ours)	**71.9**	**63.4**	**57.8**	**75.8**	**67.6**	**41.0**

Table 8: Efficiency comparison on Video-MME across video durations. Time is measured in seconds, and #Frames denotes the average number of unique frames analyzed by the VLM.

Method	Short (1–3 min)			Medium (3–30 min)			Long (30–60 min)			Overall		
	Acc.	Time (s)	#Frames	Acc.	Time (s)	#Frames	Acc.	Time (s)	#Frames	Acc.	Time (s)	#Frames
Uniform	70.2	1.1	0	57.1	1.1	0	48.2	1.1	0	58.5	1.1	0
CLIP Retrieval	68.8	3.5	0	58.6	7.5	0	45.8	32.8	0	57.7	14.6	0
VLM judgment	70.7	22.8	157.7	58.0	58.4	516.3	51.7	286.2	2465.9	60.1	122.4	1046.6
RACER	**72.6**	22.4	57.6	**60.6**	29.0	64.0	**52.7**	54.9	67.1	**61.9**	35.4	62.9

A Comparison across Question Types

To assess RACER's effectiveness across question types, Table 7 compares uniform sampling, direct CLIP selection, VLM judgment-based selection, and RACER across six Video-MME categories using LLaVA-OV-7B as the answering model. Category-level accuracy is reported over all 2700 questions. RACER achieves the highest accuracy across all six categories, demonstrating consistent benefits across diverse query requirements.

B Efficiency across Video Durations

Table 8 compares frame-selection methods on Video-MME across three video-duration groups. RACER achieves the highest accuracy in every group while scaling more efficiently than VLM judgment-based selection. As videos increase from short to long, VLM judgment-based selection analyzes 157.7 to 2465.9 distinct frames, with inference time rising from 22.8 to 286.2 seconds. RACER analyzes only 57.6–67.1 distinct frames, with runtime increasing from 22.4 to 54.9 seconds. On long videos, RACER achieves a $5.2\times$ speedup while improving accuracy from 51.7 to 52.7. Overall, RACER improves accuracy from 60.1 to 61.9 with a $3.5\times$ speedup. These results demonstrate the scalability benefits of delegating candidate-pool scoring to a retrieval tool while limiting Vid-LLM analysis to a compact set of frames.

C Additional Hyperparameter Ablation

Table 9 examines the DPP relevance parameter α, entropy sensitivity β, original-query weight w_0, and lightweight Vid-LLM generation temperature τ on Video-MME with LLaVA-OV-7B. Each sweep varies one parameter, with default settings marked by † and the best observed result in bold.

C.1 Effect of α in DPP Sampling

The parameter α balances relevance and visual diversity in DPP sampling. Accuracy increases from 60.41% at $\alpha = 0.55$ to a peak of 61.93% at $\alpha = 0.75$, while larger values yield no further improvement. The results favor balancing relevance and diversity over placing excessive emphasis on relevance.

C.2 Effect of β in Entropy-Based Fusion

The parameter β controls the influence of entropy on sub-query weights, with larger values favoring lower-entropy retrieval signals. Accuracy varies only from 61.89% to 62.11% over $\beta \in [2.0, 2.4]$.

Table 9: Hyperparameter ablation on Video-MME with LLaVA-OV-7B. Each column pair reports a separate parameter sweep. Acc. denotes accuracy (%); † marks the default setting; bold indicates the best observed result.

DPP Relevance		Entropy Sensitivity		Original-Query Weight		Generation Temperature	
α	Acc.	β	Acc.	w_0	Acc.	τ	Acc.
0.55	60.41	2.0	**62.11**	0.4	61.37	0.6	60.30
0.65	60.63	2.1	61.96	0.5	61.22	0.7	61.00
$0.75^{\dagger}$	**61.93**	$2.2^{\dagger}$	61.93	$0.6^{\dagger}$	**61.93**	$0.8^{\dagger}$	**61.93**
0.85	60.89	2.3	61.89	0.7	61.22	0.9	60.59
0.95	61.52	2.4	61.93	0.8	60.48	1.0	60.89

Table 10: Ablations of sub-query aggregation, reflection depth, and reflection input on Video-MME with LLaVA-OV-7B. Acc. denotes accuracy (%). † marks the default setting, and bold indicates the best result within each group. Each reflection step refines the sub-queries after retrieval; the default two steps yield three retrieval rounds, including the initial round.

Design	Configuration	Acc. (%)
Sub-query aggregation	Uniform weighting	61.67
	Entropy-based weighting†	**61.93**
Reflection steps	0	61.52
	1	61.89
	$2^{\dagger}$	61.93
	3	**61.96**
	4	**61.96**
Reflection input	Without reflection	61.52
	Uniform overview	61.56
	Hybrid overview (temporal-distance pruning)	61.70
	Hybrid overview (visual-similarity pruning)†	**61.93**

Although $\beta = 2.0$ yields the highest accuracy, the default $\beta = 2.2$ achieves a comparable 61.93%, indicating limited sensitivity within the tested range.

C.3 Effect of the Original-Query Weight w_0

The weight w_0 balances the original-query score against the combined sub-query scores. The default $w_0 = 0.6$ achieves the highest accuracy of 61.93%, while both lower and higher weights reduce performance. In particular, accuracy drops to 60.48% at $w_0 = 0.8$, supporting the use of sub-queries alongside the original query.

C.4 Effect of the Vid-LLM Generation Temperature

The temperature τ controls sampling during sub-query generation and refinement. The default $\tau = 0.8$ achieves the highest accuracy of 61.93%, compared with 60.30% at $\tau = 0.6$ and 60.89% at $\tau = 1.0$. The non-monotonic trend indicates that neither lower nor higher generation temperature consistently improves frame selection.

D Additional Framework Design Ablations

Table 10 evaluates RACER's design choices for sub-query aggregation, reflection depth, and reflection input construction on Video-MME with LLaVA-OV-7B.

D.1 Effect of Sub-Query Aggregation

Entropy-based weighting improves accuracy from 61.67% to 61.93% over uniform weighting, supporting adaptive sub-query weighting based on retrieval-score distributions.

D.2 Effect of the Number of Reflection Steps

Accuracy improves from 61.52% without reflection to 61.89% after one step and 61.93% after two, reaching 61.96% with three or four steps. These diminishing returns support the default two-step configuration, which captures most of the gain and requires three retrieval rounds, including the initial round.

D.3 Effect of the Reflection Input

Reflection with the uniform overview achieves 61.56% accuracy, compared with 61.52% without reflection. Hybrid overviews incorporating retrieved feedback improve accuracy to 61.70% with temporal-distance pruning and 61.93% with visual-similarity pruning. The default strategy retains all feedback frames and removes the least visually similar overview frames, supporting the use of retrieved evidence and visually aligned context for refinement.

E Additional Discussion of Related Work

We further distinguish RACER from existing frame-selection methods through task decomposition. Similarity-based methods refine sampling, temporal coverage, and diversity using query–frame similarity (Liu et al., 2025; Zhang et al., 2025c; Tang et al., 2025; Sun et al., 2025), whereas judgment-based methods rely on VLM/LLM decisions to assess relevance or select frames and temporal regions (Wang et al., 2025b; Zhang et al., 2024a; Wang et al., 2024a; Zou et al., 2026). Some methods combine these paradigms but retain VLM/LLM control over frame selection. VideoAgent (Wang et al., 2024a) combines query analysis with CLIP retrieval, with the LLM determining search regions, stopping retrieval, and adding or removing frames based on its judgments. A.I.R. (Zou et al., 2026) applies VLM relevance judgments to candidates prefiltered by CLIP. RACER instead assigns query interpretation and refinement exclusively to a lightweight Vid-LLM and delegates evidence localization entirely to an embedding-based retrieval tool. The Vid-LLM generates sub-queries without scoring frames, selecting indices, or specifying temporal regions, while the tool scores candidates and selects frames under preset budgets and reflection depth. This separation persists throughout reflection, with retrieved frames informing only sub-query refinement. RACER thus addresses the Query Comprehension Gap by making implicit evidence requirements explicit and the Interpretation–Selection Gap by delegating frame selection to the retrieval tool.