ReQuest: Rethinking-based Question-Aware Frame Selection for Long-Form Video QA

Minkuk Kim[1*], Suyong Yun[1*], Young Tae Kim[1], Jinyoung Moon[2], Jinwoo Choi[1†], and Seong Tae Kim[1†]

[1] Kyung Hee University, Republic of Korea
[2] Electronics and Telecommunications Research Institute (ETRI), Republic of Korea
{asdjklfgh97,sy9267,youngtae1216,jinwoochoi,st.kim}@khu.ac.kr,
jymoon@etri.re.kr

Abstract. Recent multimodal large language models (MLLMs) have substantially advanced video understanding, yet long-form video QA remains challenging under fixed input token budgets, where uniform sampling can be inefficient for evidence localization. We propose ReQuest , an uncertainty-driven, question-adaptive keyframe selection pipeline that aligns question intent with relevant video content through selective computation. ReQuest integrates (i) a lightweight question-aware selector distilled from MLLM-generated supervision, (ii) Re-thinking Routing that triggers additional inference only when the model is uncertain with a length-adaptive criterion, and (iii) uncertainty-guided adaptive non-maximum suppression that selects temporally diverse frames while adjusting spacing based on question difficulty. As a plug-and-play method, ReQuest improves long-video QA without modifying or fine-tuning the underlying MLLM. Experiments on Video-MME, MLVU, and LongVideoBench demonstrate consistent accuracy gains with competitive computational cost, with particularly strong improvements in medium and long video regimes. The code is available at https://github.com/ailab-kyunghee/ReQuest

Keywords: Key Frame Selection · Video Question Answering · Vision Language Model

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) [11–13, 19–21, 24, 28, 43] have significantly improved visual–language understanding and demonstrated strong potential in Video Question Answering (Video QA) [14, 15, 25,47]. However, due to the inherent limitation on input token length, these models inevitably suffer from information loss. In practice, many existing approaches prioritize high visual resolution and therefore process only a limited number of frames, which can miss critical evidence in long videos. Although recent models [2, 3] increase the maximum number of input frames, they still operate

[*] Equally contributed first authors. [†] Corresponding authors.

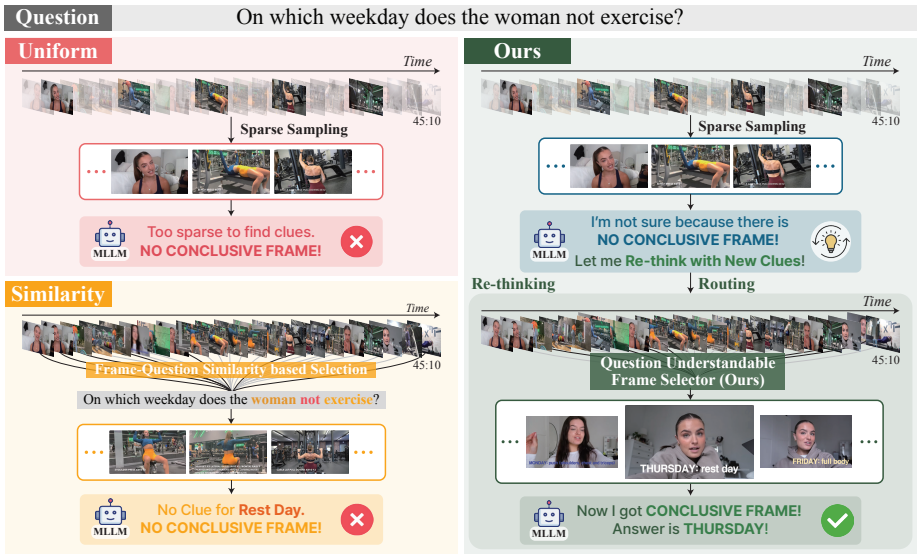


Fig. 1: Overview of selector dynamics across different video reasoning approaches. Uniform sampling selects frames at fixed intervals under a limited token budget, often missing decisive evidence. Similarity-based selectors rely solely on frame–question cosine similarity and thus struggle when key objects relevant to the answer do not explicitly appear in the question. In contrast, our question-aware keyframe selection framework performs an additional re-thinking step when no decisive evidence frame is identified, enabling global exploration of the entire video and allowing the model to uncover informative frames that existing methods overlook, ultimately yielding more reliable reasoning over long videos.

under a fixed visual token budget and must trade off frame capacity against per-frame resolution. Under this constraint, uniform sampling is widely adopted for its simplicity, yet frame allocation remains delicate: using too few frames may miss key evidence, while using many frames can introduce redundant content and low spatial detail from low resolution (Table 7).

To address this "needle-in-a-haystack" problem [37] in long videos, recent studies have proposed question-conditioned frame selection [22, 31, 32, 44]. These methods typically utilize pretrained vision–language encoders [27, 41] to compute frame–question similarity and retrieve frames that appear semantically related to the query. However, as illustrated in Fig. 1, this similarity-based design remains limited: when the key object relevant to the answer does not explicitly appear, such models lack contextual understanding of the question.

In contrast, MLLM-based selectors [10, 39] are capable of deeper semantic comprehension of the question, yet they are constrained by computational cost and thus adopt multi-stage sampling pipeline that begins with a sparse sampling stage. Consequently, they still face a risk of information loss originating from the initial sparse sampling step.

Understanding long videos demands both rich contextual reasoning and cost-efficient inference. In long-form scenarios, question tokens differ in their relevance to the final answer, meaning that simple frame–question similarity does not reliably reflect a frame's true contribution to the correct reasoning trajectory. When a question lacks explicit key objects, similarity-based approaches struggle to localize the relevant keyframes (Table 3). Furthermore, as the video length increases, uniform sparse sampling becomes even less likely to capture critical evidence frames. Therefore, long-video understanding demands a strategy that can maintain computational efficiency while performing global exploration to identify key frames that contribute most to the final answer (Table 2).

To this end, we propose a cost-efficient and question-adaptive keyframe selection framework for long-form video understanding. The design is guided by two key principles: (1) It should capture the global semantic context of the question to localize the most relevant frames. (2) It should enable efficient global exploration across the entire video within practical latency and compute limits.

Building on these insights, we introduce ReQuest , a context-aware and lightweight keyframe selector that mimics the question-understanding ability of MLLMs while enabling scalable global reasoning. In addition, we design a Re-thinking Router that dynamically determines whether additional inference is required based on the MLLM's prediction uncertainty and the question difficulty. Finally, we propose an adaptive Non-Maximum Suppression (NMS) sampling strategy that adjusts the focus of frame selection according to the model's uncertainty.

Our approach is grounded in a simple yet effective idea: leveraging the model's own responses as a self-guided signal for adaptive frame selection. This response-driven design offers an intuitive and principled way to align frame selection with the model's internal understanding of the question, enabling adaptive and scalable long-video processing. Experimental results show that our approach consistently improves performance across multiple long-video benchmarks [8, 36, 48], and that the learned selector, trained on one MLLM's responses, successfully transfers to other models [3, 15], demonstrating the generality and effectiveness of our framework. We also show that, even for high-temporal-capacity MLLMs, question-adaptive frame allocation can be more effective than uniformly denser sampling under a fixed visual-token budget, as it preserves higher per-frame visual resolution by allocating tokens to fewer but more relevant frames. Our main contributions can be summarized as:

- We design a lightweight frame selector that mimics the question-understanding ability of MLLMs to enable efficient semantic alignment between questions and frames.
- We propose a Re-thinking Routing pipeline that adaptively decides whether additional inference is necessary, guided by prediction uncertainty and question difficulty. We further introduce an adaptive NMS sampling strategy that dynamically adjusts frame spacing based on uncertainty, leading to more informative and cost-efficient keyframe selection than uniform sampling.

– We conduct extensive experiments on multiple long-video QA benchmarks, including Video-MME, MLVU, and LongVideoBench. Our method achieves state-of-the-art accuracy and demonstrates that question-adaptive frame allocation remains effective even for high-frame-capacity MLLMs under fixed practical budgets.

2 Related Work

2.1 Vision-Language Models for Video Question Answering

Video QA has advanced substantially with the development of MLLM, which jointly encodes visual and textual inputs for high-level reasoning. Early video-oriented systems such as VideoLLaMA [43], Video-ChatGPT [24], and Video-LLaVA [19] extend LLMs with visual encoders to process frame sequences, while more recent general-purpose models including Qwen-VL [1], LLaVA-OneVision [15], and InternVL [5] demonstrate strong zero-shot capabilities across both images and videos. However, handling long videos remains challenging, as the number of visual tokens grows rapidly with video duration. To address this issue, prior work has explored strategies such as frame compression, temporal subsampling, temporal grounding, and memory-based representations [9,18,28,35,40]. While these approaches help reduce computational cost, they still require selecting a limited set of frames before the MLLM can perform reasoning. Motivated by this constraint, we build our long-video understanding framework on recent general-purpose MLLMs (i.e., LLaVA-Video [45], LLaVA-OneVision [15], and Qwen3-VL [2]) and focus on designing a question-adaptive frame selection mechanism that identifies the most informative frames for effective long-video reasoning.

2.2 Query-Related Frame Selection

A large body of work has explored selecting question-relevant frames to process long videos efficiently. The most common approach estimates the relevance between a question and each frame using CLIP [27] or SigLIP [41] based similarity, and methods such as BOLT [22], MDP3 [31], Q-Frame [44], and AKS [32] extend this paradigm with strategies including importance sampling, dynamic-resolution selection, and policy-based retrieval. While computationally efficient, these similarity-driven techniques often struggle to capture semantic cues when the information required to answer the question is not visually explicit. Beyond similarity-driven selection, LVNet [26] introduces a training-free hierarchical keyframe selector to reduce redundant visual inputs for long-form VideoQA.

More recent work leverages the reasoning capabilities of MLLMs to analyze question–frame relationships more comprehensively. Approaches such as VideoAgent [33], VideoTree [34], LLoVi [42], VideoAgent [7], and SeViLA [38] use MLLMs or LLM agents to select visual evidence and several studies further explore training lightweight MLLMs to predict frame importance and select

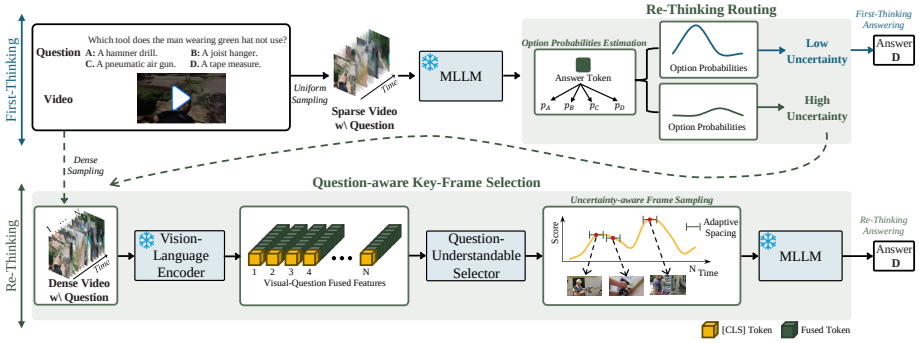


Fig. 2: Overview of the proposed framework. We address long-form video reasoning by uncertainty-guided routing and lightweight question-aware frame selection. Uniformly sampled frames are processed by an MLLM to estimate prediction entropy. The uncertainty signal determines whether the model directly outputs the answer or enters a re-thinking stage. In the re-thinking stage, a context-aware frame selector jointly encodes visual frames and question tokens to produce question-conditioned important scores. These scores are fused with the initial entropy and used by an adaptive NMS-based sampling strategy to select informative and non-redundant keyframes. The selected frames are then fed back into the MLLM for refined reasoning.

VideoQA inputs accordingly [10]. Frame-Voyager [39] constructs relative rankings of frame combinations using the prediction loss of a pretrained Video-LLM and trains a model to identify the most informative combinations for each query. Although these methods offer stronger semantic understanding, their computational cost necessitates multi-stage sampling pipelines that begin with a sparse sampling stage (e.g., 3600 → 128 → 32), resulting in an inherent trade-off between global exploration and efficiency in long-video settings.

To address these complementary strengths and limitations, we propose a question-adaptive frame selection framework that combines a lightweight, response-guided selector with an uncertainty-aware re-thinking routing strategy.

3 Method

We address long-form VQA with a question-aware keyframe selection framework that distills reasoning signals from an MLLM and executes question-conditioned evidence localization (Fig. 2). The pipeline has three parts: (§3.1) a question-aware MLLM-mimic selector, (§3.2) a response-driven re-thinking router, (§3.3) an uncertainty-based frame sampler.

3.1 Question-Aware Key Frame Selector

We train a lightweight selector that mimics the question-understanding behavior of an MLLM and estimates the contribution of each video frame to answering a

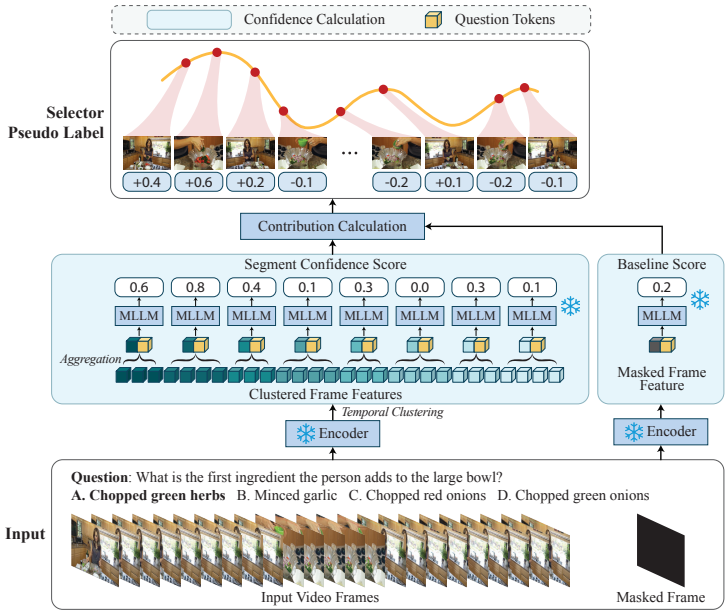


Fig. 3: Proposed pseudo-labeling pipeline. We cluster video frames into segment-level groups and query the MLLM with each segment to obtain the predicted probability of the correct answer. We then compute a baseline probability using a fully-masked visual input. By subtracting this baseline from the segment-level probability, we estimate a visual-grounded contribution score that mitigates text-prior bias. Each contribution score is used as a supervision signal for training the selector.

given question. The selector is trained using pseudo-labels extracted from the MLLM's own responses. This section describes the pseudo-label construction and the selector architecture.

Pseudo Label Generation from MLLM Responses. As illustrated in Fig. 3, for each segment s_i, we query a pretrained MLLM using s_i as the visual input and obtain multiple-choice logits $\mathbf{z}_i \in \mathbb{R}^M$ over M answer options. Let $\mathbf{p}_i = [p_{i,1}, \ldots, p_{i,M}]$, $\quad p_{i,k} = \frac{\exp(z_{i,k})}{\sum_{m=1}^{M} \exp(z_{i,m})}$, and let $g \in \{1, \ldots, M\}$ denote the index of the ground-truth answer. To define a reference for measuring contribution and reducing text-only priors, we additionally query the MLLM with a dummy zero-tensor visual input and obtain a baseline distribution $\mathbf{p}^{\text{base}} = [p_1^{\text{base}}, \ldots, p_M^{\text{base}}]$. This baseline indicates how confidently the model answers the question in the absence of visual cues. The contribution score of segment s_i is defined as

$$\Delta c_i = p_{i,g} - p_g^{\text{base}} \tag{1}$$

which measures the net visual influence of the segment. These continuous scores are used as supervision for the selector.

Selector Architecture. Each segment is first encoded by the BLIP [16] visual encoder, while the question is tokenized and processed through the BLIP text stream with cross-attention to visual tokens. Unlike dual-encoder designs, BLIP performs vision-language interaction inside the text-side fusion layers, and we use the resulting fused token representations as segment-question features. These fused features are then passed to a lightweight four-layer transformer with a learnable score token that aggregates segment-level evidence.

We first extract BLIP fused tokens at the frame level for each frame–question pair, then aggregate frames belonging to the same segment by temporal pooling to build a segment-level fused feature. Let $\mathbf{f}_i \in \mathbb{R}^{T\times D}$ denote the resulting BLIP fused token sequence for segment s_i and question q. We prepend a learnable score token to form $\tilde{\mathbf{f}}_i \in \mathbb{R}^{(T+1)\times D}$, and feed it to a lightweight transformer encoder. The final hidden state of the prepended token, $\mathbf{h}_i^{\texttt{cls}} \in \mathbb{R}^D$, is used as the aggregated representation.

The selector outputs a scalar contribution score $\hat{c}_i = \mathrm{ScoreHead}(\mathbf{h}_i^{\texttt{cls}})$, where $\mathrm{ScoreHead}(\cdot)$ is a lightweight scoring layer. The model is trained to regress toward the pseudo target Δc_i. Although training uses segment-level fused representations which reduce redundancy and enables efficient pseudo-label generation across long videos, we apply the selector at the frame level during inference. Each frame is processed with the question using BLIP to obtain a fused token sequence, and its score is predicted from the corresponding `[CLS]` representation. The selector generalizes naturally to frame-level inference because the scoring function operates on frame–question fused representations. This design enables fine-grained evidence localization while preserving the efficiency needed for global exploration in long videos.

3.2 Re-thinking Routing

Additional evidence localization is not necessary for every question. While some queries are well supported by uniformly sampled frames, others contain sparse or implicit cues that lead the MLLM to produce uncertain predictions. The router determines whether a question–video pair should follow a uniform path or activate the selector for re-thinking.

Uncertainty and Length-Aware Scoring. Given a question–video pair, we first run the MLLM on uniformly sampled frames and obtain an option probability vector $\mathbf{p} = [p_1, \ldots, p_M]$. The model's uncertainty is measured using the entropy

$$u = -\sum_{k=1}^{M} p_k \log p_k \tag{2}$$

As uncertainty naturally increases for longer videos, where uniform sampling becomes less reliable, we incorporate a length-aware correction term. Let N denote the number of frames in the video (capped at $N_{\max}$). We define a normalized length term as

$$r_{\mathrm{len}}(N) = \min\left(1,\ \frac{N}{N_{\max}}\right) \tag{3}$$

which maps N into $[0, 1]$. The effective routing threshold is then

$$\tau_{\text{eff}} = \tau_0 - \gamma_{\text{len}}\, r_{\text{len}}(N) \tag{4}$$

This yields a stricter threshold for short videos—where the model is less likely to be uncertain—and a more permissive one for long videos, enabling the router to better capture uncertainty arising from extended temporal context.

Rule-Based Routing Decision. The routing decision is entirely algorithmic and does not involve learning:

$$\text{Decision} = \begin{cases} \text{Selection Required}, & u > \tau_{\text{eff}}, \\ \text{Uniform Path}, & u \leq \tau_{\text{eff}}. \end{cases} \tag{5}$$

If the prediction is confident, the model directly answers from uniform frames. If the prediction is uncertain, the router triggers a re-thinking stage where the selector localizes evidence relevant to the question. This rule-based design ensures that computation is spent only on questions requiring deeper visual reasoning.

3.3 Uncertainty-based Frame Sampling

The selector provides contribution scores, but an additional mechanism is required to choose a fixed number of frames from long videos while balancing global coverage and local precision. We adopt a greedy NMS strategy and modulate its suppression interval using the MLLM's prediction uncertainty.

Entropy-guided spacing. Using the uncertainty score u from Eq. (2), we adapt frame spacing for sampling. High entropy indicates that the model is uncertain and potentially missing critical visual evidence, whereas low entropy implies that uniformly sampled frames already contain sufficient cues.

To adjust the density of frame selection, we define a length-aware base spacing as $b_s = \frac{N}{K}$ where N denotes the total number of video frames after 1 fps sampling, and K is the number of frames finally fed into the MLLM. We convert entropy into a scaling coefficient $s(u) \in [s_{\min}, 1]$ using a clipped and normalized form:

$$u' = \text{clip}(u;\, u_{\min}, u_{\max}) \tag{6}$$

$$w = \left(\frac{u' - u_{\min}}{u_{\max} - u_{\min}} \right)^{\rho} \tag{7}$$

$$s(u) = s_{\min} + (1 - w)(1 - s_{\min}) \tag{8}$$

When $u < u_{\min}$, we simply set $s(u) = 1$, yielding uniform-like spacing. The effective NMS interval is then

$$\delta = \text{round}(s(u) \cdot b_s) \tag{9}$$

which shrinks under high uncertainty to enable denser, more local exploration.

Adaptive Greedy NMS. Given frame scores $\{\hat{c}_i\}_{i=1}^{N}$ from the selector (here i indexes frame candidates at inference), we apply greedy NMS with the adaptive interval δ. At each step, the highest-scoring index is selected, and scores within the window $[i-\delta, i+\delta]$ are suppressed. This is repeated until K frames are chosen. High uncertainty yields small δ, allowing the sampling process to capture tightly clustered evidence, while low uncertainty recovers globally spaced uniform sampling. This rule-based sampling strategy requires no additional learning and dynamically adapts the sampling density to the question difficulty and model confidence, enabling efficient evidence retrieval across long videos.

4 Experiments

4.1 Experimental Setting

Benchmarks and Metrics. We evaluate ReQuest on three long-video understanding benchmarks: MLVU [46] (2,174 questions, 9 categories, 12-minute average), LongVideoBench [36] validation (1,337 QA pairs, 8-minute average), and Video-MME [8] without subtitles (2,700 QA pairs, 17-minute average).
Implementation Details. We use LLaVA-Video [45] for pseudo-label generation and evaluation. Frames are sampled at 1 fps and encoded with BLIP to obtain fused frame-question representations. The selector is trained with Smooth-ℓ_1 ($\beta = 0.5$) and pairwise ranking losses (α=0.3, τ=0.25, γ_{loss}=2.0), while routing remains rule-based (Sec. 3.2). For segment-level supervision, videos are clustered with FINCH [29] (second hierarchy level). We train with AdamW (batch size 256), sampling 2,048 random segment pairs per batch. Across benchmarks, we use K=32, $u_{\min}$=0.45, $u_{\max}$=1.0, $s_{\min}$=0.2, $N_{\max}$=3600, and ρ=0.07. With γ_{len}=0.3 fixed, τ_0 is set to 0.55 (Video-MME), 0.45 (MLVU), and 0.3 (LongVideoBench).
Training Dataset. We use the multiple-choice split of LLaVA-Video-178K [45]. From 808K QA pairs, we keep 169K visually grounded samples and segment each video with FINCH at 1 fps for supervision. To remove noisy cases, we retain questions whose Δc_i distribution has a clear peak (threshold 0.20). This yields 73,468 training questions and 803,956 segment-level samples.

4.2 Comparison with State-of-the-art Methods

We compare our framework against previous long-video VQA methods across Video-MME, MLVU, and LongVideoBench. As shown in Table 1, ReQuest consistently improves the base MLLM in a plug-and-play manner, without any MLLM fine-tuning. On Video-MME, applying ReQuest to LLaVA-Video improves overall accuracy from 62.6 to 65.6, with clear gains on the medium and long subsets (59.3→64.1 and 52.2→55.8), where long-range evidence aggregation is most critical. We also observe gains on MLVU and LongVideoBench (66.7→73.9 and 58.0→60.1), indicating generalization across benchmarks. Transfer to other backbones is also consistent: LLaVA-OneVision improves across benchmarks. Additionally, Qwen3-VL improves on Video-MME (70.0→71.1), MLVU (74.0→76.2),

Table 1: Comparison of our proposed ReQuest with state-of-the-art models on benchmarks. Our approach ReQuest enables MLLMs to focus on query-related frames with plug-and-play manner, improving performance across benchmarks. #Frames denotes the number of input frames used by each method. † denotes results reproduced from the official implementation in our environment. * denotes zero-shot transfer results using a selector trained with LLaVA-Video-generated supervision.

Model	LLM Size	#Frames	LVB	MLVU m-avg	Video-MME (w.o. sub.)			
					Overall	Short	Medium	Long
Video-LLaVA [19]	7B	8	39.1	47.3	39.9	45.3	38.0	36.2
VideoChat2 [17]	7B	16	–	44.5	39.5	48.3	37.0	33.2
ShareGPT4Video [4]	8B	16	–	46.4	39.9	48.3	36.3	35.0
Chat-UniVi-V1.5 [11]	7B	64	–	–	40.6	45.7	40.3	35.8
VideoLLaMA2 [6]	7B	16	–	–	47.9	56.0	45.4	42.1
TimeSuite [40]	7B	128	–	–	46.3	–	–	41.9
Frame-Voyager [39]	7B	8	–	65.6	57.5	67.3	56.3	48.9
LongVU [30]	7B	1fps	–	65.4	60.9	64.7	58.2	59.5
NVILA [23]	8B	1024	57.7	70.1	64.0	75.0	62.2	54.8
LLoVi [42]	–	–	–	55.1	54.7	62.1	53.2	48.8
VideoTree [34]	–	–	–	60.4	60.6	67.8	59.9	54.2
LLaVA-Video† [45]	7B	32	58.0	64.7	62.6	76.2	59.3	52.2
+ ReQuest	7B	32	60.1	71.7	65.6	77.0	64.1	55.8
LLaVa-OneVision† [15]	7B	32	56.6	63.1	58.7	70.3	56.6	49.2
+ ReQuest *	7B	32	60.2	68.8	60.9	71.7	58.8	52.3
Qwen3-VL† [2]	8B	512	62.7	74.0	70.0	78.6	70.1	61.2
+ ReQuest *	8B	≤ 512	**66.3**	**76.2**	**71.1**	**80.0**	**70.8**	**62.4**

Table 2: End-to-End Cost-Accuracy on Long Videos in Video-MME. Average latency and TFLOPs are measured on 900 Video-MME long-video questions. #Frames: the number of observed (decoded) frames per video. Decode: video decoding overhead, which dominates latency. ReQuest uses Re-thinking Routing. With Re-thinking Routing, dense observation is performed only for samples routed to the re-thinking stage.

Frame Selection	Answer Model	#Frames	Decode	Feature Ext.		Selection		MLLM Inf.		Total		
			Avg(s)	Avg(s)	TFLOPs	Avg(s)	TFLOPs	Avg(s)	TFLOPs	Avg(s)	TFLOPs	Long Acc.
No Selection (Baselines)												
Uniform	LLaVA-Video	32	5.4	0.5	22.2	–	–	1.6	76.6	7.5	98.8	52.2
Uniform	Qwen3-VL-8B	512	35.9	3.4	463.5	–	–	11.6	1,564.6	50.9	2028.1	61.2
Key Frame Selection (Dense Observation)												
Similarity	LLaVA-Video	1fps→32	122.3	13.1	371.9	1.3×10^{-3}	0.8×10^{-5}	1.6	76.6	137.0	448.5	54.3
Similarity + Re-thinking	LLaVA-Video	1fps→32	89.6	9.4	256.7	0.9×10^{-3}	0.6×10^{-5}	2.7	131.0	102.1	425.7	54.9
ReQuest	LLaVA-Video	1fps→32	89.6	9.4	256.7	0.3	4.2	2.7	131.0	102.3	428.7	55.8
ReQuest	Qwen3-VL-8B	1fps→$\leq$ 512	89.0	8.8	616.0	0.3	4.2	16.2	2097.0	114.1	2714.8	62.4

and LongVideoBench (62.7→66.3). These results suggest that ReQuest captures model-agnostic, question-conditioned frame importance and remains effective even for high frame-capacity MLLMs, where dense uniform sampling can be suboptimal for evidence localization under practical token constraints.

4.3 Effect of Question-Adaptive Key-Frame Selection

End-to-End Cost-Accuracy. We report end-to-end latency and TFLOPs in Table 2, including video decoding, feature extraction, key-frame selection, and

Table 3: Question understanding without key objects in VideoMME. Results are reported on LLaVA-Video with ReQuest .

Method	VQA Acc.(%)
Uniform	51.3
Feature-based Similarity	48.0
Model-based Similarity	51.0
Ours	**53.0**

Table 4: Ablation on sampling strategies in VideoMME. Across all settings, we apply the same selector and routing module, and vary only the sampling strategy to isolate its effect. Results are reported on LLaVA-Video with ReQuest .

Sampling Strategy	Overall	Short	Medium	Long
Top-k	64.0	76.7	62.3	53.0
Greedy NMS	64.7	**77.0**	63.7	53.3
Adaptive NMS	**65.6**	**77.0**	**64.1**	**55.8**

MLLM inference. With Re-thinking Routing, dense observation is performed only for samples routed to the re-thinking stage (640/900 for LLaVA-Video and 390/900 for Qwen3-VL-8B). Routing reduces cost over plain similarity: Similarity + Re-thinking lowers latency (137.0→102.1 s) and TFLOPs (448.5→425.7) while improving accuracy (54.3→54.9). ReQuest further improves accuracy to **55.8**, indicating that it maintains practical inference cost while more effectively identifying frames that are most informative for long-video understanding.

Understanding Contextual Information of the Question. To assess whether our method captures question semantics beyond object presence, we evaluate on 400 key-object-missing questions of VideoMME. Our approach focuses on contextual cues implied by the question rather than relying on visible key objects, allowing the selector to attend to informative frames even when salient objects are absent. For a fair comparison that isolates question understanding in frame selection, we apply the same top-k sampling strategy and disable routing for all methods.Feature-based Similarity computes cosine similarity using SigLIP features, whereas Model-based Similarity uses the BLIP ITM [16] score. As shown in Table 3, our method outperforms both Uniform and similarity-based baselines. These results demonstrate that our selector identifies question-relevant frames without explicit object cues, while similarity-based selection struggles when key objects are absent, due to its heavy reliance on visual cues.

4.4 Ablation Studies

Sampling Strategy Analysis. To evaluate the effectiveness of our proposed Adaptive NMS, we compare Top-k, Greedy NMS, and Adaptive NMS on VideoMME. As shown in Table 4, Greedy NMS suppresses redundancy by enforcing a fixed temporal suppression width, but it cannot adapt the suppression gap to question-specific uncertainty. By contrast, Adaptive NMS achieves the best overall accuracy and the largest gain on long videos (55.8% vs. 53.3% for Greedy NMS) by dynamically adjusting the suppression width. This supports the benefits of uncertainty-adaptive spacing for efficient long-video selection.

Ablation of Uncertainty Estimation with Router. We analyze the effect of the uncertainty-length threshold τ_{eff} on LLaVA-Video in Table 5. At τ_{eff}=0, all questions are routed to selection (2700 routed / 0 not routed) with high cost

Table 5: Analysis of router uncertainty threshold in VideoMME. #Uniform/#Selection indicate how many questions are routed to each branch. Results are reported on LLaVA-Video with ReQuest .

#Selection	#Uniform	VQA Acc. (%)
2700	0	65.3
2171	529	65.4
1782	918	65.4
1365	1335	**65.6**
1051	1649	65.0
567	2133	64.0
0	2700	62.6

Table 6: Randomness control on router-selected subsets in VideoMME. Each cell reports overall accuracy (%). Results are reported on LLaVA-Video with ReQuest .

Re-inference policy	Routing type	
	Router-Ours	Router-Oracle
Uniform (New seed)	62.8	62.7
Selection (ours)	65.6	71.4

and limited gain; at τ_{eff}=1, all questions use uniform sampling (0 routed / 2700 not routed) and accuracy drops to 62.6%. The best result appears near balanced routing (τ_{eff}=0.55, 1365 routed / 1335 not routed). A similar pattern is observed with Qwen3-VL-8B, where the router improves accuracy over the all-selection setting on Video-MME (69.9→71.1), while reducing average latency from 52.1s to 39.3s. These results show that our uncertainty-based routing effectively balances accuracy and computation, enabling a favorable cost–accuracy trade-off via simple threshold calibration.

4.5 Discussion

Controlling for Random Re-inference Effects. We control for the possibility that our improvements stem from stochasticity introduced by an additional inference pass. We perform re-inference only on the subset of questions that uniform sampling answers incorrectly, and report the overall accuracy by aggregating the originally correct predictions with the re-inferred predictions on the missed subset. As shown in Table 6, uniform sampling exhibits negligible variation across different random seeds; averaging accuracy over three runs yields nearly identical results (62.8 with Router-Ours and 62.7 with Router-Oracle). In contrast, router-guided re-inference achieves clear gains (65.6 with Router-Ours; 71.4 with Router-Oracle). This indicates that the improvements are driven by re-inference with key frames, rather than randomness from repeated inference.

Keyframe Selection in High-Temporal-Capacity Qwen3-VL. Table 7 shows that, although Qwen3-VL-8B can accept up to 2048 frames, simply increasing the number of input frames does not consistently improve VideoMME performance: accuracy peaks at 512 frames and drops at 1024/2048 frames, suggesting that overly dense inputs can introduce redundancy and dilute effective visual evidence. Motivated by this observation, we evaluate question-aware keyframe selection in this high-capacity setting. Our method, trained with LLaVA-Video-generated supervision, uses length-adaptive ratios (50/40/15% for short/medium/long videos), selecting only 40/208/369 frames on average. Despite using

Table 7: Qwen3-VL performance on VideoMME across frame budgets. We vary the number of input frames from 256 to 2048 and report accuracy and latency to examine the resolution–temporal trade-off.

Frames	Acc.				Avg(s)
	Overall	Short	Medium	Long	
2048	67.7	78.6	68.9	55.6	53.8
1024	68.7	78.6	69.0	58.6	41.6
768	69.1	78.6	69.0	59.9	36.4
512	70.0	78.6	70.1	61.2	33.2
256	69.0	78.6	67.8	60.6	24.7
40/208/369(Ours)	71.1	80.0	70.8	62.4	39.3

Table 8: Comparison of training supervision signals on VideoMME. The first two rows are training-free similarity baselines. These similarity-based selections use cosine similarity over SigLIP features or BLIP ITM scores, respectively. the remaining rows are learned selectors trained with different supervisions on SigLIP and BLIP backbones. Results are reported on LLaVA-Video with ReQuest .

Method	Selection Backbone	Training Supervision	Overall	Short	Medium	Long
Feature-based Sim	SigLIP	-	61.6	74.8	57.8	52.1
Model-based Sim	BLIP	-	65.0	76.2	64.0	54.9
Selector	SigLIP	Feature-based Sim	61.7	75.6	57.2	52.3
Selector	BLIP	Model-based Sim	63.6	76.0	61.8	53.0
Selector	SigLIP	MLLM Response (Ours)	64.7	76.2	61.7	**56.1**
Selector	BLIP	MLLM Response (Ours)	**65.6**	**77.0**	**64.1**	55.8

substantially fewer frames than the uniform 512-frame baseline, our method achieves higher accuracy across subsets with only a modest latency increase (33.2s→39.3s), while also outperforming denser uniform budgets with comparable or higher latency. Similar gains on MLVU (74.0→76.2) and LongVideoBench (62.7→66.3) in Table 1 further support the benefit of evidence-driven frame selection for high-temporal-capacity MLLMs.

Impact of Supervision Signals on Selector Effectiveness. This experiment isolates the impact of different supervision signals used to train the selector (Table 8). We compare training-free similarity baselines and learned selectors under the same pipeline, varying only pseudo-label sources (feature-based similarity, model-based similarity, or MLLM responses). Across SigLIP and BLIP selector backbones, MLLM-response supervision yields the most consistent gains, showing the importance of question-aware supervision.

Evaluation on Open-ended VQA. We further evaluate open-ended VQA to assess broader applicability. In this setting, uncertainty is quantified as the maximum entropy over generated token distributions and used to trigger the same re-thinking routing and adaptive NMS sampling. As shown in Table 9, ReQuest consistently improves open-ended VQA performance over LLaVA-

Table 9: Open-ended VQA on Video-MME. Factual accuracy is evaluated on a 0–5 scale based on the "correctness of information" criterion of Video-ChatGPT [24].

Model	Open-Ended Video-MME (w.o. sub.)			
	Overall	Short	Medium	Long
LLaVA-Video†	2.77	3.26	2.58	2.48
+ ReQuest	3.00(↑8%)	3.45(↑6%)	2.78(↑8%)	2.76(↑11%)

Table 10: Effect of prediction granularity on VQA accuracy. Results are reported on LLaVA-Video with ReQuest .

Prediction Granularity	Overall	Short	Medium	Long
Segment-level	62.9	74.8	60.4	53.3
Frame-level	**65.6**	**77.0**	**64.1**	**55.8**

Table 11: Transferability to larger-scale Qwen3-VL backbones on Video-MME. We report accuracy without subtitles.

Model	Overall	Short	Medium	Long
Qwen3-VL-8B	70.0	78.6	70.1	61.2
+ ReQuest	**71.1**	**80.0**	**70.8**	**62.4**
Qwen3-VL-32B	73.9	81.6	73.7	66.5
+ ReQuest	**75.7**	**82.8**	**76.6**	**67.8**

Video, with the largest gains on long videos. These results show applicability beyond multiple-choice settings.

Prediction Granularity. While the selector is trained at the segment level to obtain supervision efficiently, Table 10 shows that frame-level inference yields higher accuracy at test time. This is because fine-grained visual cues within a segment may be averaged out or diluted when represented as a single segment token. In contrast, evaluating frames individually allows the model to capture subtle evidence that would otherwise be lost at the segment level, leading to more precise identification of informative moments. Thus, even with segment-level training, frame-level inference proves more effective for accurate evidence. **Transferability to Larger-Scale Backbones.** As shown in Table 11, whether ReQuest remains effective when applied to a stronger MLLM. ReQuest improves Qwen3-VL-32B from 73.9 to 75.7. The gains are consistent across all video-length subsets for the 32B model, including Short (81.6→82.8), Medium (73.7→76.6), and Long (66.5→67.8). These results suggest that ReQuest remains beneficial even as the answering MLLM becomes stronger.

Qualitative Result. As illustrated in Fig. 4, we qualitatively compare uniform sampling and our question-aware keyframe selection. In this example, uniform sampling fails to capture any conclusive frame, leading to an incorrect prediction.

Fig. 4: Qualitative comparison of uniform sampling and our ReQuest selector. ReQuest selects question-relevant moments and preserves conclusive evidence frames for correct counting.

In contrast, ReQuest assigns higher importance to query-aligned frames that clearly reveal the correct count (three Spider-Men). As a result, the MLLM receives sufficient evidence and predicts the correct answer.

5 Conclusion

We presented ReQuest uncertainty-driven and question-adaptive keyframe selection pipeline for long-form video QA. ReQuest combines a lightweight question-aware selector distilled from MLLM-generated supervision, selective Re-thinking Routing based on uncertainty, and adaptive NMS sampling to avoid redundant temporal selections. Without modifying or fine-tuning the underlying MLLM, ReQuest consistently improves long-video QA accuracy on Video-MME, LongVideoBench and MLVU at competitive cost, with notable gains in medium/long regimes compared to dense uniform sampling. Additional results indicate that question-adaptive allocation can also benefit high-frame-capacity MLLMs, where dense uniform sampling can remain suboptimal for evidence localization under practical token constraints.

Acknowledgements

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) under Grant 2020-0-00004 (Development of Provisional Intelligence Based on Long-term Visual Memory Network), Grant RS-2022-II220078, Grant IITP-2026-RS-2023-00258649, and by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant RS-2024-00334321.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond (2023)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Tang, Z., Yuan, L., et al.: Sharegpt4video: Improving video understanding and generation with better captions. vol. 37, pp. 19472–19495 (2024)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
6. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
7. Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., Li, Q.: Videoagent: A memory-augmented multimodal agent for video understanding. In: European Conference on Computer Vision. pp. 75–92. Springer (2024)
8. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075 (2024)
9. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13504–13514 (2024)
10. Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B., et al.: M-llm based video frame selection for efficient video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13702–13712 (2025)
11. Jin, P., Takanobu, R., Zhang, W., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13700–13710 (2024)
12. Kim, M., Kim, H.B., Moon, J., Choi, J., Kim, S.T.: Do you remember? dense video captioning with cross-modal memory retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13894–13904 (2024)
13. Kim, M., Kim, H.B., Moon, J., Choi, J., Kim, S.T.: Hicm^2: Hierarchical compact memory modeling for dense video captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4293–4301 (2025)
14. Kim, W., Choi, C., Lee, W., Rhee, W.: An image grid can be worth a video: Zero-shot video question answering using a vlm. IEEE Access (2024)
15. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)

16. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
17. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025)
18. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2025)
19. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
20. Lin, K., Ahmed, F., Li, L., Lin, C.C., Azarnasab, E., Yang, Z., Wang, J., Liang, L., Liu, Z., Lu, Y., et al.: Mm-vid: Advancing video understanding with gpt-4v (ision). arXiv preprint arXiv:2310.19773 (2023)
21. Liu, S., Zhang, C.L., Zhao, C., Ghanem, B.: End-to-end temporal action detection with 1b parameters across 1000 frames. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18591–18601 (2024)
22. Liu, S., Zhao, C., Xu, T., Ghanem, B.: Bolt: Boost large vision-language model without training for long-form video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3318–3327 (2025)
23. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4122–4134 (2025)
24. Maaz, M., Rasheed, H., Khan, S., Khan, F.: 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). pp. 12585–12602 (2024)
25. Min, J., Buch, S., Nagrani, A., Cho, M., Schmid, C.: Morevqa: Exploring modular reasoning models for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13235–13245 (2024)
26. Park, J., Ranasinghe, K., Kahatapitiya, K., Ryu, W., Kim, D., Ryoo, M.S.: Too many frames, not all useful: Efficient strategies for long-form video qa. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3569–3588 (2026)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
28. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323 (2024)
29. Sarfraz, S., Murray, N., Sharma, V., Diba, A., Van Gool, L., Stiefelhagen, R.: Temporally-weighted hierarchical clustering for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11225–11234 (2021)
30. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024)

31. Sun, H., Lu, S., Wang, H., Chen, Q.G., Xu, Z., Luo, W., Zhang, K., Li, M.: Mdp3: A training-free approach for list-wise frame selection in video-llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24090–24101 (2025)
32. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29118–29128 (2025)
33. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)
34. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3272–3283 (2025)
35. Weng, Y., Han, M., He, H., Chang, X., Zhuang, B.: Longvlm: Efficient long video understanding via large language models. In: European Conference on Computer Vision. pp. 453–470. Springer (2024)
36. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
37. Ye, J., Wang, Z., Sun, H., Chandrasegaran, K., Durante, Z., Eyzaguirre, C., Bisk, Y., Niebles, J.C., Adeli, E., Fei-Fei, L., et al.: Re-thinking temporal search for long-form video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8579–8591 (2025)
38. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-chained image-language model for video localization and question answering. Advances in Neural Information Processing Systems **36**, 76749–76771 (2023)
39. Yu, S., Jin, C., Wang, H., Chen, Z., Jin, S., Zuo, Z., Xu, X., Sun, Z., Zhang, B., Wu, J., et al.: Frame-voyager: Learning to query frames for video large language models. arXiv preprint arXiv:2410.03226 (2024)
40. Zeng, X., Li, K., Wang, C., Li, X., Jiang, T., Yan, Z., Li, S., Shi, Y., Yue, Z., Wang, Y., et al.: Timesuite: Improving mllms for long video understanding via grounded tuning. In: International Conference on Learning Representations. vol. 2025, pp. 38057–38081 (2025)
41. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
42. Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21715–21737 (2024)
43. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
44. Zhang, S., Yang, J., Yin, J., Luo, Z., Luan, J.: Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22056–22065 (2025)
45. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
46. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025)
47. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
48. Zou, B., Yang, C., Qiao, Y., Quan, C., Zhao, Y.: Language-aware visual semantic distillation for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27113–27123 (2024)

ReQuest: Rethinking-based Question-Aware Frame Selection for Long-Form Video QA

Supplementary Material

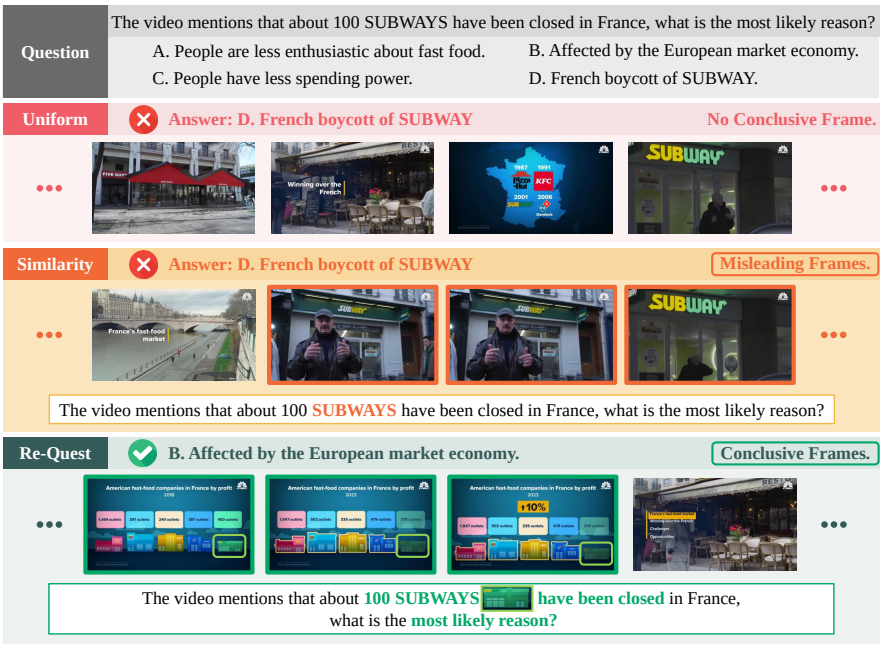


Fig. 5: Qualitative comparison of uniform sampling, cosine similarity, and our ReQuest selector. Uniform sampling misses crucial evidence due to sparse fixed-interval selection, while cosine similarity over-focuses on lexical matches such as *SUBWAY* and *France*. In contrast, ReQuest retrieves the economically informative frames needed to identify the cause of the store closures, leading to the correct answer.

A Additional Analysis of ReQuest

Qualitative Example. Figure 5 presents a qualitative comparison of uniform sampling, cosine similarity, and our ReQuest selector on a long-video question asking why about 100 SUBWAY stores were closed in France. Uniform sampling distributes frames sparsely across the video, but this fixed allocation misses the key evidence needed to infer the cause of the closures. Cosine similarity is strongly biased toward lexical cues such as *SUBWAY* and *France*, assigning high relevance to visually matched storefront scenes and logo-bearing frames, yet failing to capture the underlying semantic intent of the question. As a result,

Table 12: End-to-End Cost under Sparse Observation on Long Videos in Video-MME. Average latency and TFLOPs are measured on 900 Video-MME long-video questions. Sparse key-frame selection methods first observe 128 decoded frames and then select 8 frames for answering. Sel. Size denotes the total parameter count and memory footprint used to compute frame scores at inference, including frozen backbone components. #Frames denotes the number of observed decoded frames and selected input frames. Decode denotes the overhead for decoding the observed frames.

Frame Selection	Sel. Size	Answer Model	#Frames	Decode		Feature Ext.		Selection		MLLM Inf.		Total	
				Avg(s)	TFLOPs	Avg(s)	TFLOPs	Avg(s)	TFLOPs	Avg(s)	TFLOPs	Avg(s)	TFLOPs
No Selection (Baselines)													
Uniform	–	LLaVA-Video	32	5.4	–	0.5	22.2	–	–	1.6	76.6	7.5	98.8
Uniform	–	Qwen3-VL-8B	512	35.9	–	3.4	463.5	–	–	11.6	1,564.6	50.9	2028.1
Key Frame Selection (Sparse Observation)													
FrameVoyager [35]	1.2B / 2.2GB	LLaVa-One-Vision	128→8	8.6	–	4.1	82.2	0.1×10^{-1}	1.6	0.3	19.0	13.3	102.8
Similarity	0.2B / 0.8GB	LLaVA-Video	128→8	14.1	–	0.8	15.1	0.1×10^{-3}	0.2×10^{-8}	0.3	17.6	15.2	32.7
ReQuest	0.2B / 0.9GB	LLaVA-Video	128→8	6.0	–	0.5	11.5	0.1×10^{-1}	0.2	0.4	24.2	6.9	35.8
ReQuest	0.2B / 0.9GB	LLaVA-One-Vision	128→8	10.4	–	0.7	17.2	0.1×10^{-1}	0.2	0.5	30.1	11.6	47.5

Table 13: Ablation Study on Dense Sampling Rates in VideoMME. We vary the initial dense sampling rate used for candidate frame extraction and report average latency, TFLOPs, and long-subset accuracy. Uniform denotes the baseline without re-thinking. Results are reported on LLaVA-Video with ReQuest .

Method	Avg(s)	TFLOPs	Acc
Uniform	7.5	98.8	52.2
Ours (0.25 fps)	32.5	228.4	54.7
Ours (0.5 fps)	55.8	295.2	54.9
Ours (1 fps)	102.3	428.7	**55.8**

both methods are misled toward an incorrect answer. In contrast, our ReQuest selector successfully retrieves the economically informative frames, including statistics showing that the number of SUBWAY stores in France decreased by roughly 100 from 2018 to 2023, together with evidence that competing fast-food brands grew during the same period. Because ReQuest is trained with pseudo supervision distilled from MLLM response behaviors, it learns to focus not only on surface word overlap but also on the core meaning of the query. This example shows that ReQuest can identify semantically decisive evidence that simpler sampling and similarity-based strategies fail to retrieve.

Sparse-Observation Cost Comparison. We further report the end-to-end cost of key-frame selection methods under a sparse-observation setting in Table 12, where selection-based methods first observe 128 decoded frames and select 8 frames for answering. Compared with similarity-based selection, ReQuest substantially reduces latency on LLaVA-Video (15.2s→6.9s) with comparable TFLOPs. Under the matched LLaVA-OneVision setting, ReQuest also shows lower latency and TFLOPs than FrameVoyager (13.3s/102.8 TFLOPs → 11.6s/47.5 TFLOPs). These results show that ReQuest remains computationally competitive even when the observation budget is sparse.

Table 14: Performance of Ours on VideoMME with different number of input frames. We evaluate LLaVA-Video and our method (ReQuest) under 4, 8, 16, and 32 input frame settings, and analyze the changes in overall accuracy as well as performance across the Short, Medium, and Long subsets.

Model	#Frames	Overall	Short	Medium	Long
LLaVA-Video† [45]	4	52.8	61.8	50.2	46.2
+ **ReQuest**	4	**61.5**	**75.0**	**58.3**	**51.2**
LLaVA-Video† [45]	8	55.5	67.8	52.2	46.6
+ **ReQuest**	8	**63.0**	**76.4**	**59.8**	**52.9**
LLaVA-Video† [45]	16	60.1	71.4	58.7	50.1
+ **ReQuest**	16	**63.9**	**76.4**	**60.4**	**54.8**
LLaVA-Video† [45]	32	62.6	76.2	59.3	52.2
+ **ReQuest**	32	**65.6**	**77.0**	**64.1**	**55.8**

Robustness Across Different Dense Sampling Rates. We further evaluate the robustness of ReQuest across different dense sampling rates before frame selection by varying the initial sampling rate among 0.25 fps, 0.5 fps, and 1 fps, while keeping all sampling hyperparameters fixed (Table 13). As the sampling rate becomes lower, the overall pipeline cost decreases substantially, with average latency reduced from 102.3s at 1 fps to 55.8s at 0.5 fps and 32.5s at 0.25 fps, while TFLOPs also drop from 428.7 to 295.2 and 228.4, respectively. Despite the increasingly sparse candidate pool, ReQuest maintains clear gains over uniform sampling at all rates, achieving accuracies of 55.8, 54.9, and 54.7 at 1 fps, 0.5 fps, and 0.25 fps, respectively. Notably, all three settings surpass the uniform sampling baseline (52.2, Table 2 in the main paper), confirming that ReQuest provides consistent gains regardless of the dense sampling rate. These results suggest that ReQuest remains effective even at lower dense sampling rates, while providing a favorable efficiency-accuracy trade-off under tighter latency constraints.

Effect of Ours Across Different Frame Budgets. We analyze the effect of the input frame budget on Video-MME performance by evaluating both LLaVA-Video and ReQuest under 4, 8, 16, and 32 input frame settings (Table 14). As the number of input frames decreases, the baseline performance drops noticeably, particularly on the medium and long subsets where relevant evidence is more sparsely distributed over time. In contrast, ReQuest consistently improves accuracy across all frame budgets. Even with only 4 input frames, ReQuest improves overall accuracy by 8.7 points over the corresponding 4-frame baseline and yields clear gains across all subsets. Notably, the 4-frame variant of ReQuest outperforms the 16-frame baseline on three of four subsets, including 51.2 vs. 50.1 on the Long subset. These results suggest that selecting question-relevant frames can be more effective than simply increasing the input frame budget, and that our selector can identify a compact yet highly informative set of keyframes even under extremely limited budgets. As the frame budget increases, the improvement remains consistent, reaching 65.6 overall accuracy at 32 frames. While ReQuest improves performance across all subsets, the gains on the medium and long subsets

Table 15: Qwen3-VL performance on LongVideoBench across different frame budgets. We vary the number of uniformly sampled input frames from 256 to 2048. For ReQuest , the first-thinking stage uses the same uniformly sampled frames as each baseline to estimate prediction uncertainty and determine routing. Questions routed to re-thinking are processed by the selector, which selects approximately 275 frames for re-thinking answering. The selector and all pipeline settings for ReQuest remain identical across all rows; only the first-thinking frame budget varies.

Method	#Frames	Acc.
Qwen3-VL-8B	2048	61.9
+ **ReQuest**	275	66.0
Qwen3-VL-8B	1024	62.0
+ **ReQuest**	275	66.2
Qwen3-VL-8B	512	62.7
+ **ReQuest**	275	66.3
Qwen3-VL-8B	256	64.0
+ **ReQuest**	275	66.0

remain substantial across different frame budgets, indicating that our question-aware selector is particularly effective when reasoning depends on evidence distributed sparsely over time. Overall, these results suggest that ReQuest remains effective across different frame budgets and is especially beneficial in low-budget settings, where information loss is more severe.

Qwen3 Input Frame Budget Ablation on LongVideoBench. In the main paper (Table 7), we showed that Qwen3-VL does not benefit from simply increasing the uniform frame budget on Video-MME, and that ReQuest with fewer selected frames outperforms denser uniform inputs. Table 15 provides additional evidence of this trend on LongVideoBench. Across all uniform budgets from 256 to 2048 frames, Qwen3-VL shows inconsistent gains: accuracy does not monotonically improve with more frames, peaking at 64.0 with 256 frames and dropping to 61.9 with 2048 frames. This confirms that naively increasing the number of input frames can introduce redundancy and dilute the visual evidence available for reasoning, consistent with our observation on Video-MME. In contrast, ReQuest achieves 66.0–66.3 using only 275 selected frames across all settings, consistently outperforming every uniform baseline by a clear margin. The performance of ReQuest remains stable regardless of the initial frame budget used in the first-thinking routing stage, indicating that the selector reliably identifies informative frames even when the routing is performed on sparser inputs. Notably, the 256-frame uniform baseline uses a comparable frame budget to ReQuest (256 vs. 275 frames), yet ReQuest achieves a clear improvement (66.0 vs. 64.0), suggesting that the gains arise from question-aware selection rather than frame count reduction alone. These results further support that question-adaptive frame allocation is more effective than uniform dense sampling

Table 16: Ablation Study on Minimum Spacing Ratio s_{min} of Adaptive NMS in VideoMME. Results are reported on LLaVA-Video with ReQuest .

s_{min}	Overall	Short	Medium	Long
0.0	64.8	**77.0**	63.6	53.8
0.1	65.0	**77.0**	63.1	54.9
0.2	**65.6**	**77.0**	**64.1**	**55.8**
0.3	65.3	**77.0**	64.0	54.8
0.4	64.8	**77.0**	62.9	54.6
0.5	64.6	76.9	63.0	53.8

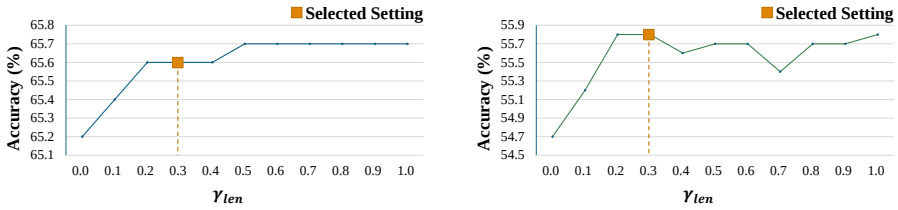
Table 17: Ablation Study on Uncertainty Sensitivity Exponent ρ of Adaptive NMS in VideoMME. Results are reported on LLaVA-Video with ReQuest .

ρ	Overall	Short	Medium	Long
0.01	64.9	**77.0**	63.6	54.1
0.02	64.9	**77.0**	63.8	54.0
0.03	64.9	**77.0**	63.7	54.1
0.04	65.0	**77.0**	63.8	54.3
0.05	65.2	**77.0**	64.0	54.7
0.06	65.3	**77.0**	63.9	55.0
0.07	**65.6**	**77.0**	64.1	**55.8**
0.08	**65.6**	**77.0**	64.1	55.7
0.09	65.5	**77.0**	**64.2**	55.2
0.10	65.5	**77.0**	64.1	55.4

for high-temporal-capacity MLLMs, and that this benefit generalizes across benchmarks.

Adaptive NMS Hyperparameter Sensitivity. We analyze the sensitivity of the entropy-guided spacing in Adaptive NMS by varying its two key hyperparameters, the minimum spacing ratio $s_{\min}$ and the uncertainty sensitivity exponent ρ, while keeping all other settings fixed. Note that $s_{\min}$ and ρ are the only tunable hyperparameters in Adaptive NMS; the remaining parameters ($u_{\min}$, $u_{\max}$) are held fixed throughout all experiments. In Table 16, we vary $s_{\min}$ while fixing $\rho = 0.07$, and in Table 17, we vary ρ while fixing $s_{\min} = 0.2$. As shown in Tables 16 and 17, ReQuest maintains strong performance across a broad range of values for both hyperparameters. Specifically, varying $s_{\min}$ from 0.0 to 0.5 yields overall accuracy in the range of 64.6-65.6, while sweeping ρ from 0.01 to 0.10 results in overall accuracy between 64.9 and 65.6. The effect of both hyperparameters is more pronounced on the medium and long subsets than on the short subset. This is expected, as our length-aware routing (Eq. 5) lowers the re-thinking threshold for longer videos, directing more long-video questions into the Adaptive NMS stage where these parameters take effect. Based on these results, we use $s_{\min} = 0.2$ and $\rho = 0.07$ as the default setting. This setting achieves the best overall performance while remaining in a stable region of the hyperparameter space. Overall, these results suggest that entropy-guided spacing is reasonably robust to moderate changes in its hyperparameters and provides an effective mechanism for balancing temporal coverage and redundancy in long-video frame selection. Notably, the same default values of $s_{\min} = 0.2$ and $\rho = 0.07$ are used consistently across all three benchmarks (Video-MME, MLVU, and LongVideoBench) as well as in the cross-model transferability experiments with LLaVA-OneVision and Qwen3-VL, without any benchmark-specific or model-specific tuning.

Sensitivity of Length-Aware Routing. We further analyze the effect of the length-aware routing weight γ_{len} on overall and long-subset accuracy (Fig. 6). While the main paper focuses on the routing threshold itself, this analysis isolates



(a) Overall accuracy **(b)** Long subset accuracy

Fig. 6: Sensitivity of the length-aware routing weight γ_{len} on VideoMME. We vary γ_{len} while fixing the base routing threshold τ_0 at 0.55. Results are reported on LLaVA-Video with ReQuest .

the effect of the length-aware routing. As shown in Fig. 6, increasing γ_{len} improves both overall accuracy in Fig. 6(a) and long video subset accuracy in Fig. 6(b) over the case without length-aware routing ($\gamma_{\text{len}} = 0.0$). This is consistent with our motivation in Sec. 3.2: as video length increases, uniform sampling becomes less reliable and the model's uncertainty naturally grows, making additional evidence localization more beneficial. The improvement is especially pronounced on the long subset, where accuracy increases from 54.7% to 55.8%, confirming that the length-aware term effectively captures the increased question difficulty associated with longer videos. Although larger values of γ_{len} yield slightly higher overall accuracy (up to 0.1%), increasing γ_{len} also routes more questions into the re-thinking stage, incurring additional computational cost. We therefore select $\gamma_{\text{len}} = 0.3$ as the default, as it achieves strong performance while maintaining computational efficiency by limiting re-thinking to the questions that benefit most from it.

B Additional Implementation Details

To improve reproducibility and provide additional details on the practical efficiency of our method, we report implementation details of our selector training, pseudo-label generation, and inference setup. The selector is trained with an initial learning rate of 1×10^{-4} using a cosine annealing learning rate scheduler. Our selector contains 13.27M parameters. Under our training setup, each epoch takes approximately 245 seconds, and the selector is trained for 20 epochs on Video-MME, 2 epochs on LongVideoBench, and 1 epoch on MLVU. Pseudo-label generation is performed offline and requires 12 hours on 64 GPUs for 803K samples. Since this cost is incurred only once during data preparation, it does not affect inference-time efficiency. For zero-shot transfer with LLaVA-OneVision, we fix $\gamma_{\text{len}} = 0.7$ and set τ_0 to 0.4 for Video-MME, 0.55 for MLVU, and 0.55 for LongVideoBench. For zero-shot transfer with Qwen3-VL, we fix $\gamma_{\text{len}} = 0.20$ and set τ_0 to 0.50 for Video-MME, and 0.15 for LongVideoBench. For MLVU,

we instead set τ_0 to 0.10 and $\gamma_{\text{len}} = 1.00$. Selector training and inference are conducted on an NVIDIA RTX A6000 GPU.