REPAIRABLE UNCERTAINTY: SELECTIVE TEMPORAL REACQUISITION FOR LONG-VIDEO UNDERSTANDING WITH MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) can process only a limited number of frames from long videos due to their large context requirements. Existing methods typically rely on uniform sampling or a separate frame selector. Uniform sampling can miss brief but decisive events, while selector-based pipelines often rely on a weaker selector than the base MLLM, potentially discarding useful evidence before it reaches the answerer. We observe that under partial evidence, an MLLM can remain uncertain about the answer while its temporal attention is concentrated on a small set of query-relevant regions. We call this repairable uncertainty: the answer remains uncertain, but focused attention provides a signal for where the missing evidence may lie. We introduce Selective Temporal Reacquisition (STR), a lightweight test-time method that uses a frozen MLLM's answer uncertainty and temporal relevance to decide whether and where to reacquire video evidence. Starting from a uniform grid of F frames, STR uses the MLLM's temporal attention to construct a new sampling density, triggering a budget-matched second pass only when the answer is uncertain and attention is sufficiently focused. We further introduce STR-Refine, which gates repeated observations using predictive uncertainty and accumulates relevance across them, and STR-Lite, which reduces the visual-token cost of reacquisition. Across five benchmarks (LongVideoBench, VideoMME, MLVU, VideoVista, and Video-MME-v2) and four MLLMs (Qwen3-VL-8B, Qwen3-VL-2B, Qwen3-VL-32B, and InternVL3-8B), STR consistently improves long-video understanding and achieves state-of-the-art performance. These results establish STR as a test-time framework for selectively reacquiring missing evidence.

1 INTRODUCTION

When humans inspect a long video under limited time, they rarely examine every moment uniformly. Even partial observations, together with prior knowledge of how events typically unfold, can suggest where to look next: after observing the beginning of an interaction, we can anticipate which moments are likely to contain the missing evidence needed to resolve an uncertainty. This is consistent with active perception, where an observer uses its current state of knowledge to determine what information to acquire next (Gibson, 1966; Bajcsy, 1988). Modern MLLMs, meanwhile, encode broad visual-semantic priors through large-scale multimodal pretraining (Bai et al., 2025; Zhu et al., 2025). We hypothesize that these priors, together with the partial evidence already observed, allow the model to infer where missing evidence relevant to its uncertainty may lie. This suggests a form of perceptual self-correction: partial evidence may be insufficient for a correct answer, yet sufficient to guide the model toward the evidence needed to correct it.

Long-video understanding makes this distinction especially important. Practical compute and token budgets make frame selection central when reasoning over recordings containing thousands of frames (Tang et al., 2025). Uniform sampling can miss a brief but decisive event. Existing frame-selection methods instead construct a more informative subset through query-conditioned ranking, relevance, or coverage (Yu et al., 2025; Tang et al., 2025), typically by scoring representations from a candidate pool before forming the final visual context. Moreover, this selection is often delegated to a lightweight encoder or smaller model whose visual-semantic reasoning is more limited than that

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.

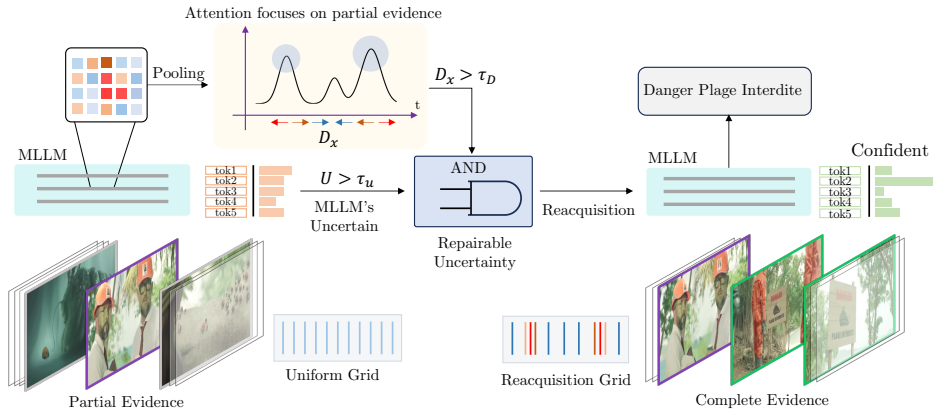


Figure 1: **Selective Temporal Reacquisition (STR).** With uniform sampling, there is only partial evidence, i.e., a group of workers. In this scenario, the MLLM is uncertain about the answer, but attention focuses on partial evidence. **STR** uses this discrepancy, selectively, to acquire new query-relevant evidence and thereby improve the MLLM's long video understanding.

of the downstream MLLM, creating an information bottleneck in which potentially useful evidence may be discarded before the answerer ever observes it. Expanding the candidate pool reduces this risk but increases processing cost, while restricting it can omit the decisive evidence altogether. This motivates two questions: *when an MLLM gives a wrong answer, can the same inference still indicate where to look to correct it? And can the frozen answerer direct this reacquisition from only 64 uniformly sampled frames, without a separate frame-selection model or scoring a larger candidate pool?*

Our premise is that a capable MLLM can answer effectively when sufficient visual evidence is available, while a sparse uniform view may expose only partial evidence. Importantly, an uncertain answer does not imply that the observation is uninformative: we observe that the model's attention can remain concentrated on query-relevant regions even when its answer is uncertain. *Partial evidence may therefore be insufficient to answer correctly, yet sufficient to indicate where additional evidence should be acquired.* We call this opportunity **repairable uncertainty**. Building on this observation, we introduce **Selective Temporal Reacquisition (STR)**, a bounded self-correction policy that uses the frozen answerer's own first-pass signals to determine both *whether* and *where* to reacquire visual evidence. STR redistributes the same frame budget toward temporally relevant regions while retaining broad video coverage, requiring neither MLLM finetuning nor a separate temporal-selection model.

Unlike response-level self-refinement, which uses model-generated feedback to revise an output (Madaan et al., 2023), STR uses the answer attempt to revise the *observation itself*. It triggers reacquisition only when the answer remains uncertain and the attention-derived proposal meaningfully redistributes temporal coverage, while keeping the answerer frozen and the per-observation frame budget fixed. This turns the initial inference into both a diagnostic of whether more evidence is needed and a guide for where to acquire it. We further extend STR with STR-Refine, which enables iterative, uncertainty-gated reacquisition, and STR-Lite, which improves efficiency by reducing the visual-token cost of reacquired views.

In summary, our key contributions are: **1)** We characterize *repairable uncertainty*: an unsuccessful answer pass can still provide temporal guidance toward missing evidence. **2)** We propose *STR*, a lightweight test-time self-correction policy that uses answer uncertainty to determine *whether* to reacquire and the frozen answerer's temporal attention to determine *where*, requiring neither MLLM finetuning nor a separate temporal-selection model. **3)** We extend STR with STR-Refine, which accumulates temporal relevance across repeated observations, and STR-Lite, which reduces the visual-token cost of reacquired views. **4)** We validate selective reacquisition through controlled evidence-acquisition analysis and natural-video benchmarks. Under matched stopping rules, we compare targeted and uniform acquisition using temporal F1, and we evaluate efficiency using measured inference latency in addition to token and FLOP counts (Wen et al., 2025).

2 Related Work

Adaptive evidence acquisition. Video evidence acquisition has been explored through iterative retrieval, hierarchical expansion, and temporal zooming (Wang et al., 2024; 2025; Ding et al., 2026), as well as keyframe localization before answering (Yu et al., 2023). Contemporary methods (Li et al., 2026b; Wang et al., 2026; Yang et al., 2026) incorporate attention and uncertainty into this process: DAFS selects frames using a smaller MLLM's attention (Wang et al., 2026), while ReQuest invokes a trained question-aware selector when the initial answer is uncertain (Kim et al., 2026). CADER routes uncertain answers through a crop–verify loop with a separate verifier (Yang et al., 2026), whereas EcoFrame combines output entropy and target-model attention to expand a candidate pool and reselect frames with a relevance encoder as its frame budget grows (Li et al., 2026b). These approaches retain separate candidate-scoring or verification stages, whose computational cost is not reflected in the answerer's final frame budget. In contrast, STR exploits *repairable uncertainty*: partial evidence can leave the answer uncertain while the frozen answerer's temporal attention indicates where missing evidence may lie. Rather than scoring a larger candidate pool, STR converts the initial answer pass's cached temporal readout directly into a new sampling density. In base STR, answer uncertainty and proposed support deformation jointly gate targeted reacquisition at a fixed per-observation frame budget. The initial answer attempt thus supplies the signals for both whether and where to acquire further evidence, without training a temporal selector or invoking a separate selection or verification model.

Query-conditioned frame selection. Query-conditioned frame selection constructs a compact visual context by evaluating candidate frames before answering. Learned selectors train on MLLM-ranked frame combinations (Yu et al., 2025), predict both frame relevance and count (Li et al., 2026a), or optimize selection policies through reinforcement learning (Lee et al., 2026; Zhou et al., 2026). Training-free alternatives balance query relevance with diversity or temporal coverage (Sun et al., 2025; Tang et al., 2025), while WFS-SB allocates frames across clips separated by wavelet-detected semantic boundaries (Chen et al., 2026). BOLT uses inverse-transform sampling from encoder-derived query–frame relevance (Liu et al., 2025). The contemporary method DAFS scores candidates using a smaller MLLM's attention (Wang et al., 2026), whereas Temporal Chain of Thought prompts the answering VLM itself to select frames (Arnab et al., 2025). Despite these differences in supervision and scoring, constructing the answerer's context still requires candidate evaluation or explicit frame-selection calls. STR instead converts the frozen answerer's temporal readout from a sparse initial answer pass directly into a sampling density for subsequent observations. Partial evidence thus guides the acquisition of additional frames, without first scoring a larger candidate pool, training a temporal selector, or generating a separate frame-selection response.

Visual-token compression. Visual-token compression reduces the cost of sampled frames using shallow-layer decoder attention (Chen et al., 2024), embedding diversity (Alvar et al., 2025), token merging (Bolya et al., 2023; Shang et al., 2025), or joint spatio-temporal redundancy reduction (Du et al., 2026; Zhang et al., 2026c). FlexSelect uses reference-layer attention to rank visual tokens and trains a lightweight query-conditioned selector to reproduce these rankings (Zhang et al., 2025b). These methods compress already acquired frames, whereas reacquisition changes which frames are observed. STR-Lite combines both: it keeps the initial pass dense and prunes only reacquired views, without changing their sampled frame locations. Beyond distilling backbone attention, its router aligns its temporal distribution with the reacquisition proposal, coupling token allocation to regions selected for further observation.

3 STR: Repairing the Temporal Observation

Overview. STR uses the frozen MLLM's uncertainty and temporal guidance to decide whether and where to reacquire evidence. Starting from a uniform view, it redistributes the same frame budget while retaining broad coverage. STR-Refine accumulates relevance across gated observations; STR-Lite reduces the token cost of reacquisition.

3.1 Uncertainty and Repairability

Given a video V with n_x frames and a question q, let $x = (V, q)$. The frozen MLLM answers from $O_0 = V[\mathcal{S}_0]$, where $\mathcal{S}_0$ is an ordered uniform grid of F frame indices. We index sampled frames by $f = 0, \ldots, F-1$ and their M spatial tokens by $i = 1, \ldots, M$.

Predictive uncertainty. We define predictive uncertainty as the mean full-vocabulary entropy over autoregressively generated answer tokens (Malinin & Gales, 2021). For the generated answer $\widehat{y}_0 = (\widehat{y}_{0,1}, \ldots, \widehat{y}_{0,L})$, let $p_m(\omega) = p_\theta(\omega \mid O_0, q, \widehat{y}_{0,<m}), \omega \in \mathcal{V}$, denote the model's next-token distribution at answer position m. We define

$$U(x) = -\frac{1}{L} \sum_{m=1}^{L} \sum_{\omega \in \mathcal{V}} p_m(\omega) \log p_m(\omega). \tag{1}$$

These distributions are produced during the normal answer generation pass, so computing U requires no additional model calls.

Temporal guidance. We use *temporal guidance* to quantify whether the current attention readout provides a meaningful direction for where to look next. Let $\widetilde{a}_{f,i}$ denote the normalized attention weight of spatial token i in sampled frame f. We obtain frame-level relevance by pooling over spatial tokens:

$$p_f = \sum_{i=1}^{M} \widetilde{a}_{f,i}, \qquad \sum_{f=0}^{F-1} p_f = 1. \tag{2}$$

The attention extraction and normalization procedure is described in Appendix B. A nearly uniform temporal profile provides little direction for reacquisition, whereas a structured profile shifts sampling toward selected temporal regions. We measure the strength of this guidance by how much the resulting proposal $\mathcal{S}_{\mathrm{re}}(x)$ redistributes the initial uniform grid $\mathcal{S}_0$. For frame position f, define the normalized displacement

$$\delta_f = \frac{(\mathcal{S}_{\mathrm{re}})_f - (\mathcal{S}_0)_f}{n_x - 1}, \qquad \bar{\delta} = \frac{1}{F} \sum_{f=0}^{F-1} \delta_f.$$

The deformation score is

$$D(x) = \frac{1}{\sqrt{F}} \left\| \boldsymbol{\delta} - \bar{\delta} \mathbf{1} \right\|_2. \tag{3}$$

Mean-centering isolates redistribution of temporal resolution from a common shift of the grid. STR uses $D(x) \geq \tau_D$ as its criterion for sufficient temporal guidance.

Repairable and non-actionable uncertainty. Predictive uncertainty and temporal guidance can coexist: the answer may be uncertain while the readout still indicates where to reacquire evidence. We call this *repairable uncertainty*. Conversely, if the answer is uncertain but the readout provides insufficient temporal guidance, we retain the initial answer:

$$\widehat{y}_{\mathrm{STR}}(x) = \begin{cases} \widehat{y}_{\mathrm{re}}(x), & U(x) > \tau_U,\ D(x) \geq \tau_D \quad \text{(repairable)}, \\ \widehat{y}_0(x), & U(x) > \tau_U,\ D(x) < \tau_D \quad \text{(non-actionable)}, \\ \widehat{y}_0(x), & U(x) \leq \tau_U \quad \text{(confident)}. \end{cases} \tag{4}$$

Here $\widehat{y}_{\mathrm{re}}$ is the answer obtained from $V[\mathcal{S}_{\mathrm{re}}]$ using the same frozen MLLM, and both thresholds are fixed. Thus, STR reacquires evidence only when uncertainty about *what to answer* coexists with guidance on *where to look next*.

3.2 Selective Temporal Reacquisition (STR)

Given the temporal relevance profile p from eq. (2), we redistribute the frame budget toward relevant regions while preserving broad temporal coverage.

For the initial grid $\mathcal{S}_0$, let $t_0 < \cdots < t_{F-1}$ denote the sampled frames' normalized positions in the video. Define Voronoi-cell boundaries $e_0 = 0$, $e_F = 1$, and $e_f = (t_{f-1} + t_f)/2$ for $f = 1, \ldots, F-1$, with widths $w_f = e_{f+1} - e_f$. We distribute each relevance mass p_f uniformly within its cell and mix it with a uniform temporal floor:

$$\mu_f = (1 - \eta) p_f + \eta w_f, \qquad \eta \in (0, 1). \tag{5}$$

This defines density μ_f / w_f within each cell and a continuous, piecewise-linear CDF Φ. We place the next F frames at its midpoint quantiles:

$$\begin{aligned} t_f^{\mathrm{re}} &= \Phi^{-1}\left(\frac{f + 1/2}{F} \right), \qquad f = 0, \ldots, F-1, \\ \mathcal{S}_{\mathrm{re}}(x) &= \mathrm{Proj}_{n_x, F}\left((t_f^{\mathrm{re}})_{f=0}^{F-1} \right). \end{aligned} \tag{6}$$

Here $\mathrm{Proj}_{n_x,F}$ maps the desired positions to F distinct native frame indices in chronological order ($n_x \geq F$). The resulting view concentrates frames in relevant temporal regions while the uniform floor maintains broad coverage.

3.3 STR-REFINE

STR-Refine extends STR to iterative reacquisition, using predictive uncertainty to decide when to stop and accumulated temporal relevance to determine where to look next. Starting from the uniform view $\mathcal{S}_0$, the frozen MLLM answers from $O_j = V[\mathcal{S}_j]$ at observation j, producing $\widehat{y}_j$. We allow at most J observations, with $J = 4$ in our experiments.

When another observation is required, let $p_{j,f}$, $e_{j,f}$, and $w_{j,f}$ denote the frame relevance, cell boundaries, and cell widths of view j. Its relevance density on the video timeline is $\rho_j(t) = p_{j,f}/w_{j,f}$ for $t \in [e_{j,f}, e_{j,f+1})$. We initialize $\bar{\rho}_0 = \rho_0$ and accumulate subsequent readouts as

$$\bar{\rho}_j(t) = (1 - \beta_{\text{acc}})\bar{\rho}_{j-1}(t) + \beta_{\text{acc}}\rho_j(t), \qquad j \geq 1, \tag{7}$$

where $\beta_{\text{acc}} \in (1/2, 1)$ is fixed, emphasizing the newest readout while retaining earlier relevance. We apply the uniform floor after accumulation and sample the next F frames from resulting CDF:

$$\begin{aligned} \Phi_j(t) &= \int_0^t \left[(1-\eta)\bar{\rho}_j(t') + \eta\right] \mathrm{d}t', \\ \mathcal{S}_{j+1} &= \mathrm{Proj}_{n_x,F}\left(\left(\Phi_j^{-1}\left(\tfrac{f+1/2}{F}\right)\right)_{f=0}^{F-1}\right). \end{aligned} \tag{8}$$

Let $U_j(x)$ denote Eq. 1 evaluated for the answer $\widehat{y}_j$. After each observation, $U_j(x) \leq \tau_U^{\text{exit}}$ returns $\widehat{y}_j$ immediately. If the observation budget is exhausted, we return the lower-uncertainty of the final two answers:

$$\widehat{y}_{\text{Refine}}(x) = \begin{cases} \widehat{y}_{J-2}, & U_{J-2}(x) < U_{J-1}(x), \\ \widehat{y}_{J-1}, & \text{otherwise.} \end{cases} \tag{9}$$

STR-Refine-Lite follows the same procedure while keeping the initial view dense and pruning only reacquired views; accumulated relevance is computed from each actual sparse answer pass. Thus, a confident first answer requires only one model call, and four observations are a maximum rather than a fixed execution cost.

3.4 STR-LITE

STR-Lite reduces the visual-token cost of reacquired views while leaving STR's temporal support unchanged. The initial view remains dense; a learned query-conditioned router prunes only refinement views while maintaining coverage across frames. Given the language sequence and visual features of a refinement view, the router produces frame logits $s \in \mathbb{R}^F$ and spatial-token logits $u \in \mathbb{R}^{F \times M}$. We define $P = \mathrm{softmax}(\beta s)$, $A_f = \mathrm{softmax}(u_{f,\cdot}/\gamma_{\text{patch}})$, where $\beta > 0$ and $\gamma_{\text{patch}} > 0$, and score each visual token by

$$\pi_{f,i} = P_f A_{f,i}. \tag{10}$$

Given retention ratio $\lambda_{\text{keep}} \in (0, 1]$, we retain $K = \mathrm{round}(\lambda_{\text{keep}} FM)$ tokens, with a per-frame minimum

$$k_{\min} = \lfloor \min(K/F, \kappa M) \rfloor, \tag{11}$$

where $\kappa \in [1/M, 1]$ controls coverage. We first retain each frame's $k_{\min}$ highest-scoring tokens, then fill the remaining budget using the largest $\pi_{f,i}$ values globally.

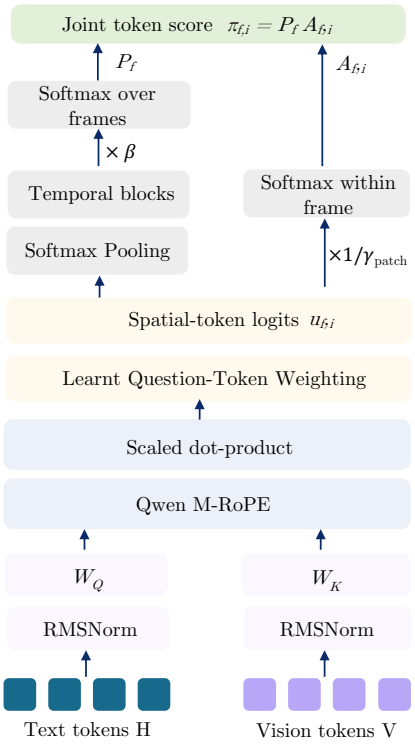


Figure 2: STR-Lite architecture.

Label-free router training. Both training stages use the frozen backbone's layer-$\ell^\star$ attention as a teacher. Let B_f be the normalized spatial attention distribution within frame f, and define the temporally sharpened target $Q_f = \frac{p_f^{1/\gamma_{\text{teach}}}}{\sum_{g=0}^{F-1} p_g^{1/\gamma_{\text{teach}}}}$, $\gamma_{\text{teach}} > 0$.



Figure 3: **Partial evidence guides STR-Refine toward complete visual evidence.** Rows show successive observations. Purple-bordered frames contain partial evidence; green-bordered frames contain complete evidence. C denotes answer confidence. The temporal relevance profiles (black curves) highlight query-relevant regions even before the decisive details are visible, guiding subsequent reacquisition. *Left:* STR-Refine recovers the brief event following the described interaction in the flood footage. *Right:* it captures the four dispersed actions to establish their temporal order

Stage one distills the teacher's spatial and temporal relevance. With $P^\circ = \mathrm{softmax}(s)$ and $A_f^\circ = \mathrm{softmax}(u_{f,\cdot})$, we optimize

$$\mathcal{L}_1 = \frac{1}{F}\sum_{f=0}^{F-1} D_{\mathrm{KL}}(A_f^\circ \| B_f) + D_{\mathrm{KL}}(P^\circ \| Q). \tag{12}$$

Stage two aligns the router's temporal allocation with the reacquisition proposal μ from Eq. 5. Using $\widetilde{\mu} = (1-\varepsilon_{\mathrm{KL}})\mu + \frac{\varepsilon_{\mathrm{KL}}}{F}\mathbf{1}$, we optimize

$$\mathcal{L}_2 = D_{\mathrm{KL}}(P \| \widetilde{\mu}). \tag{13}$$

Both stages use reverse KL, encouraging retained probability mass to concentrate on teacher-supported regions, while the per-frame minimum in Eq. 11 preserves temporal coverage. Router architecture, attention normalization, optimization, and checkpoint selection are detailed in Appendix A.3.

4 Experimental Results

We first compare STR and STR-Lite with the baselines across benchmarks and video durations. We then analyze when reacquisition helps, how its components contribute, and what can be retained across repeated observations and changing questions.

4.1 Benchmarks and Baselines

Benchmarks. We evaluate on five benchmarks. LongVideoBench (Wu et al., 2024) tests referring reasoning over long videos. VideoMME (Fu et al., 2025) spans short, medium and long video understanding. MLVU (Zhou et al., 2025) covers diverse long-video tasks and genres. VideoVista (Li et al., 2024b) tests video perception and reasoning. Video-MME-v2 (Fu et al., 2026) tests aggregation, temporal modeling and reasoning. We also present a streaming question-answering application

Table 1: Accuracy (%) with Qwen3-VL-8B and a 64-frame per-view budget. All STR variants use gating; Refine variants use at most four observations. VMME and VMME-v2 denote VideoMME and Video-MME-v2 (Fu et al., 2026). Δ gives STR-Refine's gain over the Qwen3-VL baseline.

	LongVideoBench				VMME				MLVU	VideoVista	VMME-v2
Method	**All**	**Long**	**Med.**	**Short**	**All**	**Long**	**Med.**	**Short**			
Qwen3-VL	63.0	53.4	62.9	78.1	67.4	56.0	66.6	79.7	67.6	75.3	29.3
AKS (Tang et al., 2025)	64.8	57.4	63.1	78.1	67.8	58.4	66.1	78.8	71.1	75.9	30.7
BOLT (Liu et al., 2025)	63.6	56.2	63.6	75.1	68.1	59.7	66.7	78.1	72.7	73.7	31.1
OneClip-RAG (Chen et al., 2025)	63.4	54.3	62.4	78.7	66.8	56.8	65.0	78.7	68.2	74.7	28.8
Q-Frame (Zhang et al., 2025a)	61.9	50.9	62.6	78.1	67.1	57.7	64.4	79.3	67.4	75.1	28.7
AdaQ (Zhang et al., 2026b)	64.9	58.7	65.1	74.5	67.7	57.7	66.3	79.1	71.8	75.0	31.6
CADER (Yang et al., 2026)	61.9	53.4	65.1	71.5	67.0	54.9	67.1	79.1	68.2	75.0	30.7
STR	66.2	56.9	67.5	79.2	69.4	58.4	68.8	**81.2**	70.8	76.3	31.3
STR-Lite *(ours)*	66.3	58.2	65.8	**79.5**	68.5	58.0	67.9	79.9	69.7	76.0	30.5
STR-Refine *(ours)*	**69.0**	**62.4**	**69.4**	78.7	70.1	59.3	**71.1**	79.9	74.3	**78.7**	**33.8**
STR-Refine-Lite *(ours)*	67.1	58.9	68.9	77.8	**70.3**	**60.6**	70.6	79.8	72.1	78.6	33.5
Δ Accuracy	+6.0	+9.0	+6.5	+0.6	+2.7	+3.3	+4.5	+0.2	+6.7	+3.4	+4.5

on StreamingBench SQA (Lin et al., 2026) in Appendix B. Evidence localization uses T*'s annotated LongVideoBench subset (Ye et al., 2025). Dataset protocols are in Appendix A.1.

Baselines. AKS (Tang et al., 2025) balances query relevance with temporal coverage. BOLT (Liu et al., 2025) uses inverse-transform sampling of query–frame relevance. Q-Frame (Zhang et al., 2025a) adapts frame selection and resolution to the query. OneClip-RAG (Chen et al., 2025) retrieves query-relevant clips in one step. AdaQ (Zhang et al., 2026b) adapts a quasi-Gaussian frame-sampling distribution. CADER (Yang et al., 2026) uses confidence-gated cropping and verification. Our controls include Uniform, which samples evenly spaced frames in one pass. Unconditional reacquisition always takes a second view. Random routing selects questions for reacquisition at a matched call budget. Uncertainty-gated uniform resampling requests new views when uncertain.

Implementation details. We use frozen Qwen3-VL-8B-Instruct (Bai et al., 2025) with $F{=}64$ frames per view and bfloat16 inference. Uniform views use interval midpoints; reacquisition selects distinct native frames without temporal resampling. STR is training-free. STR-Lite trains only its router and retains 60% of refinement tokens, keeping the first pass dense. Both Refine variants gate each additional observation using predictive uncertainty, up to a maximum of four observations with the same per-view frame budget. Section 4.3 reports the runtime comparison. Gate thresholds, readout settings, and training protocols are specified in Appendix A.

4.2 Quantitative results

STR outperforms state-of-the-art temporal-selection methods. Table 1 compares temporal selection and evidence acquisition using the same Qwen3-VL-8B answerer. STR-Refine improves over uniform sampling across all benchmarks in the table, with gains of 6.0 points on LongVideoBench. It also outperforms the reported temporal-selection baselines on LongVideoBench, while STR-Refine-Lite achieves the highest reported VideoMME accuracy. These gains extend to MLVU without retuning the accumulation rule, supporting observation repair as a broadly applicable approach for improving frozen MLLMs.

Gains are largest on medium and long videos. At a fixed frame budget, longer videos leave larger gaps between sampled observations. Consistent with this limitation, STR-Refine gains most on the medium and long duration groups in Table 1. On LongVideoBench, the improvement reaches 9.0 points for long videos, compared with 0.6 for short videos. VideoMME shows the same broad pattern. Targeted reacquisition is most useful when the initial view provides sparse temporal coverage.

Accumulating evidence across regions improves compositional reasoning. Some questions require complementary observations from different parts of a video. STR-Refine retains earlier relevance while incorporating newly highlighted regions, allowing each view to cover multiple temporal peaks. This is useful for counting and ordering actions in MLVU (Zhou et al., 2025), and for relating events and ordering scenes in LongVideoBench (Wu et al., 2024). Relative to uniform sampling,

the gains reach 10.7 points on MLVU *Action Count* and 11.7 points on LongVideoBench *Event before/after Event*, with improvements on both ordering tasks. These results support reacquiring complementary evidence for reasoning across time.

STR transfers across MLLM families and sizes without training. STR-Refine reuses each backbone's own attention readout with the same uncertainty gate, evidence accumulation, and observation cap. Alongside the Qwen3-VL-8B results, gains on Qwen3-VL-2B, Qwen3-VL-32B, and InternVL3-8B (Zhu et al., 2025) support transfer across model families and sizes (Table 2). The smaller Qwen3-VL-2B model also benefits, suggesting that even a less capable model's attention can guide itself.

Table 2: Plug-and-play transfer across backbones. LongVideoBench accuracy (%).

Backbone	Uniform	STR-Refine	Δ
Qwen3-VL-2B	55.72	58.64	+2.92
InternVL3-8B	62.68	65.74	+3.07
Qwen3-VL-32B	63.28	69.48	+6.20

4.3 Efficiency

Efficient execution. Unlike frame-selection pipelines that first score a separate candidate pool, STR derives where to look next directly from the answerer's first pass. We reconstruct temporal relevance from a cached decoder layer while leaving answer generation on its standard optimized attention path. To reduce this readout overhead, we implement a custom Triton (Tillet et al., 2019) kernel that reconstructs only the required query rows, normalizes over causally visible keys, and directly reduces attention over visual tokens without materializing the full attention matrix. The resulting reacquisition proposal is available before final answer uncertainty is known. We therefore begin decoding and preprocessing the proposed frames while the remaining decoder layers finish the first answer. The uncertainty gate then determines whether the second pass is needed; confident requests discard the prepared refinement. For requests that do refine, STR-Lite further reduces cost by pruning visual tokens only in the reacquired view, while keeping the initial pass dense and preserving the same temporal support. Thus, STR reuses the first answer pass to guide reacquisition, overlaps evidence preparation with inference, and reduces refinement-time computation.

Latency–accuracy trade-off. Figure 4 compares selective reacquisition with frame-selection baselines on VideoMME. We measure end-to-end request latency on the same 30-question panel using one A100 and batch size one, with a one-pass reference measured in each run. The main comparison keeps first-pass CPU preparation and GPU kernels fixed across methods; both two-pass variants use the same optimized CPU preparation only for refinement. Full timing and baseline implementation details are provided in Appendix C.

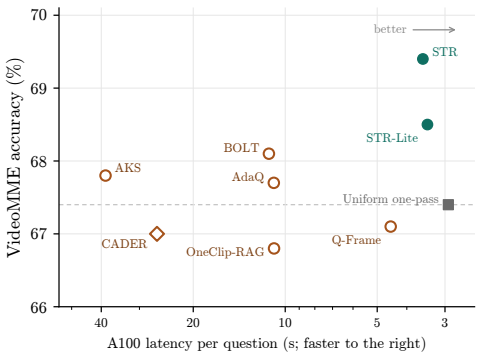
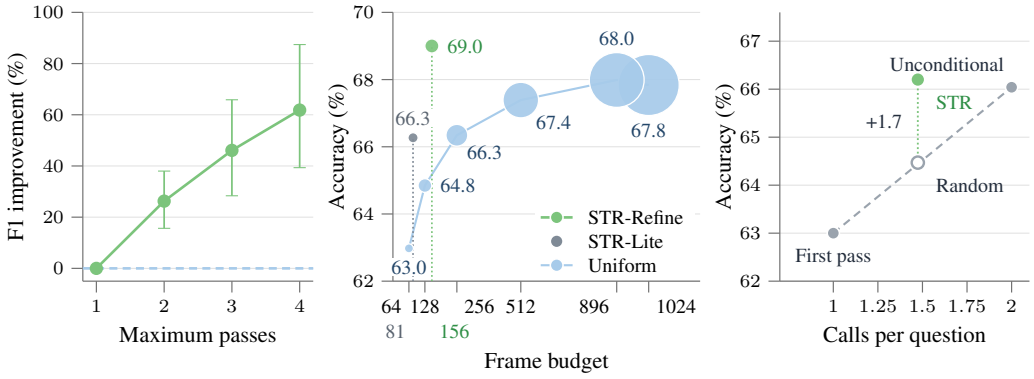


Figure 4: **Latency–accuracy trade-off on VideoMME.** Mean request latency (log scale, faster to the right) versus full-benchmark accuracy.

STR and STR-Lite improve accuracy at 1.17–1.21× one-pass latency while acquiring a second view for only about one third of questions (Figure 4). STR-Lite further reduces latency by lowering the visual-token workload of the refinement pass. In contrast, frame-selection methods incur additional candidate decoding and scoring before constructing the final observation. Under matched A100 measurements, AdaQ costs 3.5× the one-pass reference, while the other released selectors range from 1.5× (Q-Frame, 128 Long-CLIP (Zhang et al., 2024) candidates) to 12.1× (AKS, BLIP at 1 fps) per question. Even when selector computation is amortized over VideoMME's three questions per video, these costs range from 1.1–5.0× (Appendix C).

For settings that allow additional test-time computation, STR-Refine and STR-Refine-Lite extend the same selective execution to repeated observations. They reach 70.1% and 70.3% VideoMME accuracy, respectively (Table 1), while using at most four observations and requesting another view only while predictive uncertainty remains above the stopping threshold. Detailed workload, memory, and profiling results are provided in Appendix C.



(a) Evidence acquisition (b) Accuracy and compute (c) Selective reacquisition

Figure 5: (a) Relative improvement in mean temporal F1 of STR-Refine over uncertainty-gated uniform resampling on the evidence subset; error bars are 95% paired video-cluster intervals. (b) Accuracy versus frame budget; larger circles mark more model-core compute. Uniform positions are frame caps; STR positions mean frame equivalents. (c) Accuracy against calls per question. The dashed line is the random-routing expectation and the open marker its value at STR's budget.

4.4 Ablations and Analysis

Targeted reacquisition improves temporal F1 under a shared uncertainty gate. We compare attention-guided and uniform acquisition on the LongVideoBench temporal-evidence subset constructed with T*'s released protocol (Ye et al., 2025). Both methods start from the same uniform view and use the same predictive-uncertainty rule to request further 64-frame observations, up to four passes. STR-Refine places new frames using accumulated attention; the control resamples uniformly across time. We evaluate temporal F1, which combines the concentration of sampled frames near annotated evidence with coverage of the reference keyframes. Each point scores the latest acquired view, with earlier stopped views carried forward. At the four-pass cap, STR-Refine improves mean temporal F1 by 61.8% relative to uniform resampling while using slightly fewer observations. The paired confidence interval excludes zero, supporting more effective evidence acquisition under the shared stopping rule (Figure 5a). The annotations are used only for evaluation.

Selective reacquisition preserves most of the accuracy gain and recalls 87.9% of repairable errors. We test whether STR's uncertainty and attention-derived deformation gate directs additional answer calls toward questions that benefit from another observation (Figure 5c). On LongVideoBench, STR eliminates approximately 53% of second-pass calls with an accuracy loss below 0.3 percentage points relative to unconditional reacquisition. STR also outperforms random routing at the same call budget, indicating that the gate directs additional computation toward more useful reacquisitions. Repair recall makes this targeting explicit: of the first-pass errors that the recorded second pass corrects, the gate selects 87.9%, compared with 47.4% expected from random routing at the same call budget. Selective execution thus preserves most of the accuracy benefit of reacquisition while avoiding the majority of second-pass calls. The first-pass and gated STR accuracies follow the final results in Table 1. The unconditional control uses the recorded second pass from the gating analysis, including outcomes for questions that the gate skipped.

Targeted reacquisition improves the frame–compute trade-off. Figure 5b increases the uniform frame budget while preserving spatial resolution. Uniform sampling at 896 frames requires approximately 15× the model-core compute of gated STR-Refine, yet remains less accurate. STR-Lite likewise reaches comparable accuracy to a wider uniform view with substantially less computation. Selective short observations therefore offer a more favorable accuracy–compute trade-off than extending a single uniform sequence.

5 Conclusion

STR enables a frozen MLLM to use its own uncertainty and temporal attention to decide when and where to reacquire missing video evidence. Across long-video benchmarks, selective reacquisition consistently improves over uniform sampling and competitive temporal-selection methods while avoiding unnecessary additional observations. These results establish observation repair as an effective test-time mechanism for improving long-video understanding.

AI USE STATEMENT

We used generative AI tools solely to assist with grammar correction and to improve the clarity, readability, and presentation of author-written text. The tools were not used to generate or modify research ideas, methodology, experiments, results, or scientific claims, and were instructed to preserve technical terminology, method names, notation, and other domain-specific content. All AI-assisted edits were reviewed and approved by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We will release a complete reference implementation of Selective Temporal Reacquisition (STR) under an open-source license, including: (1) implementations of STR, STR-Refine, STR-Lite, and STR-Refine-Lite, including attention readout, uncertainty gating, temporal reacquisition, and iterative relevance accumulation; (2) training code and configurations for the STR-Lite router; (3) evaluation scripts and exact configurations for the five benchmarks (LongVideoBench, VideoMME, MLVU, VideoVista, and Video-MME-v2), plus the separate StreamingBench application in Appendix B; and (4) scripts for the evidence-acquisition, ablation, and efficiency analyses reported in the paper. We will also release the trained STR-Lite router checkpoints used in our experiments. Detailed dataset protocols, inference settings, hyperparameters, attention-readout implementation, and router-training procedure are provided in Appendices A–B.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal Chain of Thought: Long-video understanding by thinking in frames. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8), 1988. doi: 10.1109/5.5968.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token Merging: Your ViT but faster. In *International Conference on Learning Representations (ICLR)*, 2023.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.html.

João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the Kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. URL https://arxiv.org/abs/1907.06987.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision (ECCV)*, 2024.

Tao Chen, Shaobo Ju, Qiong Wu, Chenxin Fang, Kun Zhang, Jun Peng, Hui Li, Yiyi Zhou, and Rongrong Ji. Towards effective long video understanding of multimodal large language models via one-shot clip retrieval. *arXiv preprint arXiv:2512.08410*, 2025.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based frame selection by detecting semantic boundary for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. VideoZoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

Junhao Du, Jialong Xue, Anqi Li, Jincheng Dai, and Guo Lu. Unified spatiotemporal token compression for Video-LLMs at ultra-low retention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

James J. Gibson. *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston, 1966.

Minkuk Kim, Suyong Yun, Young Tae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. ReQuest: Rethinking-based question-aware frame selection for long-form video QA. In *European Conference on Computer Vision (ECCV)*, 2026.

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*, 2026.

Chaoyu Li, Tianzhi Li, Fei Tao, Zhenyu Zhao, Ziqian Wu, Maozheng Zhao, Juntong Song, Cheng Niu, and Pooyan Fazli. FrameOracle: Learning what to see and how much to see in videos. In *International Conference on Machine Learning (ICML)*, 2026a.

Ke Li, Jiayu Chen, Maoliang Li, Zihao Zheng, Hailong Zou, Hengyi Zhang, Xuanzhe Liu, and Xiang Chen. When and where to look: Adaptive visual evidence scheduling for efficient long video understanding. *arXiv preprint arXiv:2608.03918*, 2026b.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. VideoVista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024b.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Thomas Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2005-173.pdf.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *Transactions on Machine Learning Research (TMLR)*, 2026.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, pp. 510–526, 2016. URL https://arxiv.org/abs/1604.01753.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP3: A training-free approach for list-wise frame selection in Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Philippe Tillet, H. T. Kung, and David Cox. Triton: An intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages (MAPL)*, 2019.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, 2024.

Yilin Wang, Xiangxi Zheng, Dongxing Mao, Linjie Li, Zhengyuan Yang, Ping Yu, Rui Yan, Yuan Yao, and Alex Jinpeng Wang. Efficient frame selection for long videos at test time with attention-based MLLM selectors. *arXiv preprint arXiv:2607.15689*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Jinlong Yang, Wenhao Zhang, Kuanwei Lin, and Sijie Cheng. CADER: Confidence-aware dynamic evidence reasoning for long-video understanding. *arXiv preprint arXiv:2607.24582*, 2026.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems (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 *International Conference on Learning Representations (ICLR)*, 2025.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *European Conference on Computer Vision (ECCV)*, 2024.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. VScan: Rethinking visual token reduction for efficient large vision-language models. *Transactions on Machine Learning Research (TMLR)*, 2026a.

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*, 2026b.

Mengjie Zhang, Qihui Zhu, Tao Zhang, Shuangwu Chen, Huihuang Qin, Yu Guo, Shenghao Ye, Zijian Wen, Yunpeng Hou, Dong Jin, Xiaobin Tan, Huasen He, and Jian Yang. GSTEP: Global spatio-temporal density-driven visual token pruning for efficient video large language models. In *ACM International Conference on Multimedia (ACM MM)*, 2026c.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. FlexSelect: Flexible token selection for efficient long video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yuan Zhou, Litao Hua, Shilong Jin, Wentao Huang, and Haoran Duan. ReaSon: Reinforced causal search with information bottleneck for video understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=QmZKc7UZCy.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.