ACTIVE VISUAL PERCEPTION AND REASONING FOR LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

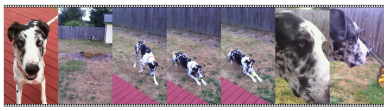
Long-video understanding requires identifying sparse, query-relevant evidence from a large amount of visually redundant content, making exhaustive video processing both inefficient and unnecessary. We present QSTAR, a Query-dependent Spatial-Temporal Active Reasoning framework for efficient long-video understanding without retraining. QSTAR casts video understanding as an iterative process of active perception, in which the model progressively concentrates its limited visual budget on moments that are most informative for answering the query. Starting from sparsely sampled frames, QSTAR exploits the VLM's internal visual reasoning signals to estimate query-dependent frame saliency and propagates these observations over the video timeline to refine a temporal sampling distribution. Successive rounds adaptively update the sampling distribution until the change between consecutive distributions falls below a threshold, after which a small set of informative frames is used for final multimodal reasoning. We further explore complementary sampling strategies and diversity-aware model fusion to improve robustness when different perceptual cues provide complementary evidence. We provide finite-round guarantees on reconstruction and induced sampling-distribution error. Extensive experiments on five video-understanding benchmarks, against 17 representative open- and closed-source VLMs and 10 methods specialized for long-video understanding, show that QSTAR consistently improves reasoning accuracy at substantially lower end-to-end latency. These results demonstrate that actively deciding where to look, conditioned on the query and the model's evolving reasoning state, provides an effective alternative to exhaustive long-video processing. Code is available at `https://anonymous.4open.science/r/QSTAR`.

1 INTRODUCTION

Recent vision–language models (VLMs) have demonstrated remarkable zero-shot and instruction-following capabilities on a growing range of video understanding tasks Team et al. (2024); Zhang et al. (2024c; 2025a); Wang et al. (2025a); Zhang et al. (2024b); Li et al. (2024c;a;b); Xu et al. (2025a). However, *long-video understanding* (LVU) remains an open challenge for a number of reasons. First, recent LVU research has focused on frame sampling techniques to address the inherent limitations of conventional exhaustive frame-by-frame processing for Video QA, which is computationally prohibitive for long videos. Second, real-world video queries often involve a small sub-sequence of frames in long videos lasting minutes to hours. Thus, static sparse sampling over long videos often risks missing critical visual cues in answering real world queries. Third, seeking query relevant visual information without frame-by-frame examination of per-frame-caption is like finding the needle in a haystack. Modern long-form benchmarks now systematically expose this difficulty by requiring accurate localization *plus* multi-hop reasoning over sparse, query-specific evidence in minute- to hour-scale videos Mangalam et al. (2023); Zhou et al. (2024); Wu et al. (2024); Fu et al. (2025). Existing approaches that rely on training query-agnostic caption to perceive query-dependent video information tend to fail miserably in locating relevant temporal and spatial visual evidence. A detailed overview on related work is provided in the next Section. A fundamental obstacle is the *sampling dilemma* Qu et al. (2025): low-density sampling risks missing critical evidence, whereas high-density sampling is both redundant and computationally prohibitive due to the quadratic scaling of transformer attention over long token sequences Ye et al. (2025). Since most VLMs operate under a strict visual-token budget, practical systems must select a small set of informative frames before the

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.

(a) Video question



Why did the dog bend down when it reached the fence in the middle of the video?

A. react to the lady B. chase ball C. play with toy **D. enjoy the ball** E. kiss it

(b) Model comparison

Model	Answer	Reasoning used for the decision	Latency (s)
ChatGPT 5.6-Sol	B	Follows or reaches the thrown ball near the fence.	60
Gemini 3.1-Pro	B	Chases the thrown ball, then bends to pick it up.	18
Qwen 3.6-Plus	B	Interprets the action as retrieving the ball during fetch.	20
QSTAR Ours	D	**Uses later frames (22.022–32.532 s), where the dog remains relaxed with the ball.**	9

Figure 1: Responses from QSTAR and three strong close-sourced VLMs on a NeXT-QA example.

final reasoning step. Thus, a fundamental research question is *how to perform query-adaptive frame selection that is semantically grounded in both spatial and temporal visual relevance and yet low latency and resource-efficient.*

We propose **QSTAR**, a **Q**uery-dependent **S**patial-**T**emporal **A**ctive **R**easoning framework for efficient long video understanding. Rather than to rely solely on an external VLM as the retrieval model, QSTAR iteratively selects a small set of informative frame samples to compute their saliency scores and measures how strongly the visual tokens support the generated answer tokens by leveraging the VLM's own *internal activation signals*. Concretely, for each meaningful answer token, we project visual-token hidden states onto the corresponding LM-head direction and subtract context-induced interference from prompt and previously generated tokens (Section 3.2). The resulting per-frame usefulness scores drive an *adaptive temporal sampling distribution* that focuses subsequent iterations on more informative moments. To densify scores over the full timeline of a long video while keeping the selection of sampled frames small, we fit a lightweight one-dimensional Gaussian process over the cached frame scores Williams & Rasmussen (1995) and update the sampling distribution until stabilization (Section 3.1). QSTAR is an active evidence-seeking agent framework that treats each LVU query as an iterative adaptive perception process without retraining the underlying VLM, and progressively refines saliency-based frame sampling using the model's internal reasoning signals. This progressive adaptive perception process enables QSTAR to refine the sampling distribution through iterative saliency estimation and resampling.

Furthermore, we optionally incorporate complementary CLIP-based sampling and model branches to provide additional evidence when different perceptual cues are beneficial. The action-fusion phase consolidates candidate outputs across branches through conflict resolution and consensus building. For example, a CLIP-based branch can complement the core saliency-based branch when the question contains a detailed description of the target frame(s). This fusion is optional, and the main-paper results use only the core saliency-based branch unless otherwise stated. QSTAR also achieves a favorable accuracy–latency trade-off compared with representative VLMs and LVU methods. Figure 1 shows an example on NeXT-QA Xiao et al. (2021), where QSTAR produces the correct answer with $2\times$, $2.2\times$, and $6.76\times$ lower latency than the three compared closed-source VLMs.

This paper makes three contributions. First, we introduce an *activation-based frame scoring* mechanism that scores answer-relevant visual tokens via answer-conditioned VLM-head projections, yielding a fine-grained query-relevant frame signal grounded in saliency-based spatio-temporal cues. Second, we develop a *training-free iterative sampling* procedure that updates a query-specific frame sampling distribution under a fixed per-round frame budget. It combines inverse-transform sampling with Gaussian-process interpolation to propagate observed saliency scores across the video timeline and guide later rounds toward frames with higher estimated relevance. Refinement terminates when the KL divergence between consecutive sampling distributions falls below a fixed threshold, serving as an inter-round change criterion. We further provide theoretical analysis of the GPR reconstruction and induced sampling distribution under standard assumptions. Third, we explore optional action fusion with complementary sampling and model branches to improve robustness when different perceptual cues provide complementary evidence. Unless otherwise stated, all main-paper results use only the core saliency-based branch, while complementary branches and fusion are evaluated separately in the Appendix. Extensive experiments across *five video QA benchmarks* compare QSTAR

with 17 representative VLMs and 10 recent long-video selection and agentic reasoning methods. QSTAR consistently improves its backbone performance while achieving substantially lower end-to-end inference latency than representative baselines at the evaluated operating points.

2 Related Work

Long-video VLMs under limited visual budgets. General-purpose VLMs and instruction-tuning recipes have improved video understanding performance Team et al. (2024); Zhang et al. (2024c;b); Li et al. (2024a;b); Zhang et al. (2025a); Wang et al. (2025a). However, benchmarks ranging from MvBench to long-form suites such as EgoSchema, MLVU, LongVideoBench, and Video-MME show that long-video QA remains difficult when the decisive evidence is sparse, subtle, and temporally dispersed Li et al. (2024c); Mangalam et al. (2023); Zhou et al. (2024); Wu et al. (2024); Fu et al. (2025); Feng et al. (2025); Qu et al. (2025); Ye et al. (2025). One response is to scale long-context video models directly, e.g., TimeChat Ren et al. (2024), Video-XL Shu et al. (2025), and LongVILA Chen et al. (2025b). Another is to reduce the per-frame cost through token compression or pruning, including LLaMA-VID, LongVU, plug-and-play token pruning, dynamic token merging, and compact key-frame representations Li et al. (2024d); Shen et al. (2024); Chen et al. (2024); Zhang et al. (2025c); Bao et al. (2025). These directions make longer inputs tractable, but they do not by themselves solve the central question of *which* moments should receive the limited visual-token budget.

Retrieval and query-adaptive frame selection. An alternative thread of research has focused on selecting frames or segments before the final reasoning step for LVU tasks. Retrieval-augmented pipelines convert long videos into captions, documents, or structured memories and then retrieve relevant evidence Luo et al. (2026); Ma et al. (2025). Similarity-based methods instead score candidate frames against the query using lightweight encoders such as CLIP Radford et al. (2021); Liu et al. (2025); Tang et al. (2025); Hu et al. (2025). More recent selectors strengthen this paradigm with explicit coverage or list-wise objectives, for example by balancing relevance and coverage or by modeling diversity and sequentiality in the selected frame set Tang et al. (2025); Sun et al. (2025). While efficient and often training-free, these methods usually rely on external similarity signals that can be brittle when the query doesn't describe the event in detail, or when there is need of temporal or compositional reasoning Qu et al. (2025); Ye et al. (2025); Feng et al. (2025).

Iterative and agentic long-video reasoning. A more adaptive alternative is *agentic* or *iterative* long-video reasoning, where a VLM iteratively explores the video timeline using tools, such as temporal localization, memory, and re-planning Fan et al. (2024); Shang et al. (2024); Chen et al. (2025a); Yang et al. (2025); Suo et al. (2025); Zhang et al. (2025b); Ventura et al. (2025); Xu et al. (2025b); Pang & Wang (2025). Recent methods Zou et al. (2026); Ben-Ami et al. (2026); Xu et al. (2026); Bai et al. (2026); Wang et al. (2025b); Zhu et al. (2026) further improve query adaptivity through iterative semantic validation, hierarchical decomposition, or task-optimized selection policies. A.I.R. Zou et al. (2026) performs deep semantic validation over adaptive micro-batches. HiMu Ben-Ami et al. (2026) decomposes queries into atomic predicates routed to lightweight unimodal experts. Multi-agent deliberation architectures Xu et al. (2026) parallelize the perceptual workload across ensembles; and reward-optimized selectors such as HORNet Bai et al. (2026) train policies that minimize downstream prediction loss. These methods often gain accuracy by paying for excessive VLM calls, auxiliary experts, or separately trained selection modules.

Our method targets the middle ground between cheap but shallow retrieval and heavy agentic exploration: Unlike existing retrieval adaptive frame selection utilizing external relevance scores, QSTAR leverages the deployed VLM's own internal activations to compute a frame-utility signal based on saliency-scores, while remaining training-free. Unlike existing agentic methods, QSTAR progressively refines the frame sampling distribution using lightweight interpolation and saliency-based and activation-grounded spatio-temporal cues, yielding a more semantically grounded selector at substantially lower cost than exploration-heavy multi-round search.

3 QSTAR Methodology

QSTAR is designed as a three-phase spatio-temporal adaptive video perception framework, aiming for efficient long video understanding. It takes a video and a question as an input, and produces the final answer through the *saliency-based perception sampling* phase and the *KL-divergence-based*

active learning phase with iterative reflection and refinement until the frame sampling distribution stabilizes. Then it enters the *action fusion* phase to generate the final answer.

We denote a video by $V = \{f_1, f_2, \ldots, f_F\}$, where F is the total number of frames. The question, together with optional answer choices and subtitles, is tokenized into $X = (x_1, \ldots, x_{n_p})$, where n_p is the number of text tokens. Each VLM call uses at most K sampled frames ($K << F$). After visual encoding, each frame contributes P visual tokens, so one VLM call contains $n_v = KP$ visual tokens in total. Let d be the hidden size of the language model. The final linear prediction layer of the language model is the *LM head*, denoted by $W \in \mathbb{R}^{|\mathcal{Y}| \times d}$, where $|\mathcal{Y}|$ is the vocabulary size. For any token $y \in \mathcal{Y}$, let $w(y) \in \mathbb{R}^d$ denote the row of W corresponding to token y. Given a hidden state $h \in \mathbb{R}^d$, the LM head produces token logits by Wh. Intuitively, $w(y)$ is the direction in hidden-state space that increases the logit of token y.

In contrast to existing agentic methods with heavy exploration, QSTAR starts to interact with a VLM using a small selection of frame samples initially drawn from a uniform distribution, and progressively refines the frame sampling distribution using saliency-based and activation-grounded spatio-temporal cues with lightweight interpolation, while remaining training-free. It yields a more semantically grounded selector than existing adaptive frame selection approaches that are based on external relevance scores, at substantially lower cost than exploration-heavy multi-round search. Figure 2 illustrates the three phase workflow of QSTAR.

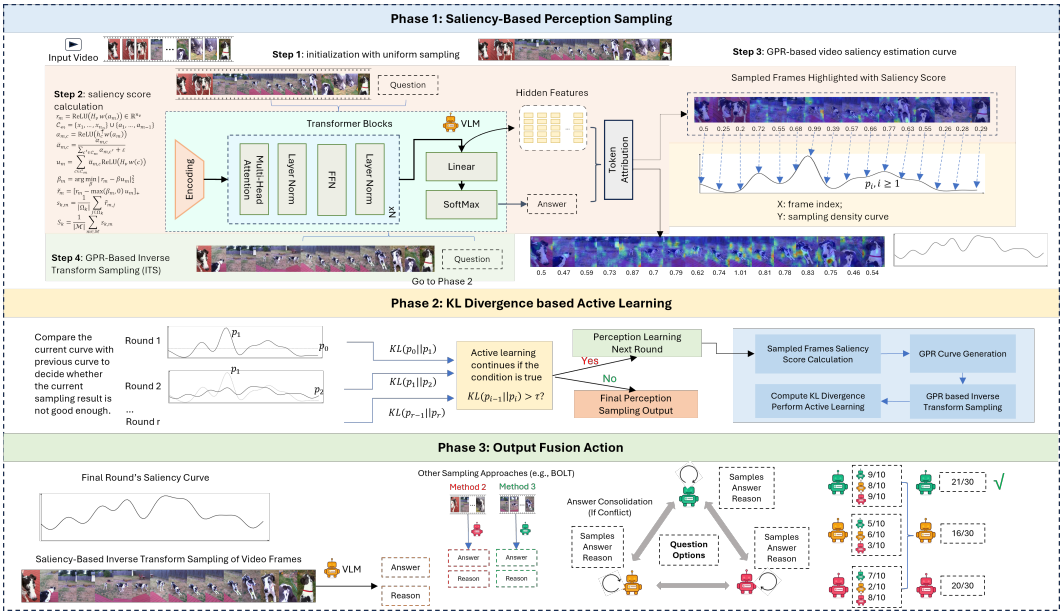


Figure 2: QSTAR system architecture with saliency-based perception sampling, KL-guided adaptive refinement, and action fusion.

3.1 Phase 1: Saliency-Based Perception Sampling

We perform the following four basic tasks in Phase 1, corresponding to the saliency-based perception loop in Figure 2. At round r, the method maintains a sampling distribution $p^{(r)} \in \mathbb{R}^F$ such that $\sum_{j=1}^{F} p_j^{(r)} = 1$ over the frame set V.

Step 1: Initialization with uniform sampling. We initialize the perception process with a uniform prior over frame positions $p_j^{(0)} = \frac{1}{F}$ where $j = 1, \ldots, F$. This step matches the initialization stage in the architecture figure and ensures that the first round does not assume any prior temporal preference.

Step 2: Sampled-frame saliency score calculation. Given the current distribution $p^{(r-1)}$, we select K frame positions using inverse transform sampling without replacement (ITS). Let $C_j = \sum_{\ell=1}^{j} p_\ell^{(r-1)}$ for $j = 1, \ldots, F$ be the cumulative distribution function. We map fixed quantiles to frame indices by binary search on $\{C_j\}_{j=1}^{F}$ and apply a deterministic uniqueness correction so that the sampled set contains exactly K distinct indices whenever possible. Denote the sampled positions

by $\mathcal{J}^{(r)} = \mathrm{ITS}(p^{(r-1)}, K)$ and the corresponding sampled frames by $\mathcal{F}^{(r)} = \{g_j \mid j \in \mathcal{J}^{(r)}\}$. We feed $\mathcal{F}^{(r)}$ and the question prompt X to the VLM: $(A^{(r)}, \mathcal{H}^{(r)}) = \mathrm{VLM}(\mathcal{F}^{(r)}, X)$, where $A^{(r)}$ is the generated answer and $\mathcal{H}^{(r)}$ contains the internal hidden states used for saliency computation. From $\mathcal{H}^{(r)}$, we compute one saliency score for each sampled frame and denote it as $S_1^{(r)}, \ldots, S_K^{(r)}$, using the activation-based scoring procedure in Section 3.2.

Step 3: Gaussian process regressor (GPR) based dense saliency curve estimation. The sampled-frame saliency scores are cached at their corresponding frame positions. Let $\tilde{s}_j^{(r)}$ denote the known score of frame g_j if it has already been sampled in any previous round. To estimate saliency values at unsampled frames, we fit a one-dimensional Gaussian process regressor (GPR) on temporal position and all cached scores accumulated so far. This yields a dense saliency curve $\hat{s}^{(r)} = \left(\hat{s}_1^{(r)}, \ldots, \hat{s}_F^{(r)}\right)$, where $\hat{s}_j^{(r)}$ is the predicted saliency of candidate frame g_j. This dense curve is the GPR-based video saliency estimation curve shown in the top right of Figure 2. Detailed process and parameters are specified in Appendix B

Step 4: GPR-based inverse transform sampling. We convert the dense saliency curve into the next-round sampling distribution by $p_j^{(r)} = \frac{\max(\hat{s}_j^{(r)}, \epsilon)}{\sum_{\ell=1}^{F} \max(\hat{s}_\ell^{(r)}, \epsilon)}$. ϵ is set as 10^{-12} in our implementation. Thus, those frames with larger predicted usefulness receive larger probability mass. The updated distribution is then used by ITS to produce/select the next-round of K frame samples. To determine whether the K sampled frames drawn from the recently updated frame sampling distribution meets the system-defined stability threshold, we evaluate the updated sampling distribution through a KL-Divergence based active learning. Before we introduce the active learning in Phase 2, we first describe how we compute the saliency score for each sampled frame in Step 2 of Phase 1.

3.2 Activation-based Saliency Score Calculation

Final-layer hidden states. After the sampled frames and text prompt are processed by the VLM, the multimodal transformer produces a hidden state for every token at every layer. We use the hidden states from the *final transformer layer*, i.e., the last contextual token representations before the LM head converts them into vocabulary logits. These states already contain the integrated visual and textual reasoning state of the model. Concretely, we extract three groups of final-layer hidden states for each VLM call:

$$H_v \in \mathbb{R}^{n_v \times d}, \qquad H_p \in \mathbb{R}^{n_p \times d}, \qquad H_a \in \mathbb{R}^{n_a \times d}.$$

Here H_v, H_p and H_a contain the final-layer hidden states at all visual-token positions, retained prompt tokens and generated answer tokens collected during greedy decoding, respectively.

Selecting meaningful answer tokens. Not every generated token is equally informative. Punctuation, special tokens and some function words often carry little visual information. We keep a subset of meaningful answer-token positions $\mathcal{M} \subseteq \{1, \ldots, n_a\}$ and compute saliency only from these tokens.

Raw visual evidence for one answer token. For each selected answer token position $m \in \mathcal{M}$, let a_m be the corresponding generated token. We measure how strongly each visual token supports this answer token by projecting the visual-token hidden states onto the LM-head direction of a_m, which is $\mathrm{ReLU}(H_v\, w(a_m)) \in \mathbb{R}^{n_v}$. The jth entry of r_m is large when the jth visual token is strongly aligned with the logit direction that increases the probability of token a_m.

Reducing context interference. Some activation may be explained by prompt text or by previously generated answer tokens, rather than by truly informative visual evidence. To reduce this effect, we define the context set for the mth answer token as $\mathcal{C}_m = \{x_1, \ldots, x_{n_p}\} \cup \{a_1, \ldots, a_{m-1}\}$. For each context token $c \in \mathcal{C}_m$, let $h_c \in \mathbb{R}^d$ be its final-layer hidden state taken from H_p or H_a. We compute its relevance to the current answer token as $\alpha_{m,c} = \mathrm{ReLU}\big(h_c^\top w(a_m)\big)$. After normalization $\bar{\alpha}_{m,c} = \frac{\alpha_{m,c}}{\sum_{c' \in \mathcal{C}_m} \alpha_{m,c'} + \varepsilon}$, we form a context-induced visual map $u_m = \sum_{c \in \mathcal{C}_m} \bar{\alpha}_{m,c}\, \mathrm{ReLU}(H_v\, w(c))$.

We then subtract the portion of the raw map that can be explained by this context map: $\beta_m = \arg\min_\beta \|r_m - \beta u_m\|_2^2$, $\tilde{r}_m = [r_m - \max(\beta_m, 0) u_m]_+$. This step suppresses visual evidence that is already predictable from surrounding text context and preserves evidence that is more specific to the current answer token. Theorem A.6 characterizes this residualization.

From token saliency to frame saliency. The visual tokens are ordered frame by frame. Let $\Omega_k \subseteq \{1, \ldots, n_v\}$ be the visual-token indices belonging to the kth sampled frame. For answer token m, we aggregate its cleaned visual evidence on sampled frame k by $s_{k,m} = \frac{1}{|\Omega_k|} \sum_{j \in \Omega_k} \tilde{r}_{m,j}$. In implementation, the token scores of each frame are first reshaped to a 2D patch grid, filtered with a small rank-Gaussian filter, and then averaged spatially. The final saliency score of the kth sampled frame is the average over all selected answer tokens: $S_k = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} s_{k,m}$. These frame scores supervise the saliency curve update in Phase 1.

3.3 Phase 2: KL-Divergence-Based Active Learning

In Phase 2, we introduce KL-divergence based active learning to measure the change from the previous sampling distribution and determine whether another perception sampling round should be performed. Concretely, upon obtaining the updated sampling distribution $p^{(r)}$, we compare it with the previous-round distribution $p^{(r-1)}$ using KL divergence $D_{\mathrm{KL}}\big(p^{(r)} \,\|\, p^{(r-1)}\big)$. If $D_{\mathrm{KL}}\big(p^{(r)} \,\|\, p^{(r-1)}\big) > \tau$, it indicates that the current saliency-based GPR curve may still change substantially, so another round of saliency-based perception sampling should be performed progressively. Here τ is a QSTAR system supplied hyperparameter. Only when the KL-divergence of the current distribution against the previous distribution is below τ, the latest distribution update is regarded as sufficiently small and the adaptive refinement process stops. Upon the stabilization, we sample the final K frames from the learned distribution: $\mathcal{F}^* = \{g_j \mid j \in \mathrm{ITS}(p^{(r)}, K)\}$, and run the VLM once more to obtain the final output of the saliency branch as $\hat{y}^{(0)} = \mathrm{VLMAnswer}(\mathcal{F}^*, X)$. The key idea is that QSTAR does not rely on a fixed external relevance model. Instead, it progressively learns a query-specific spatio-temporal sampling distribution from the VLM's own internal saliency-based reasoning signal.

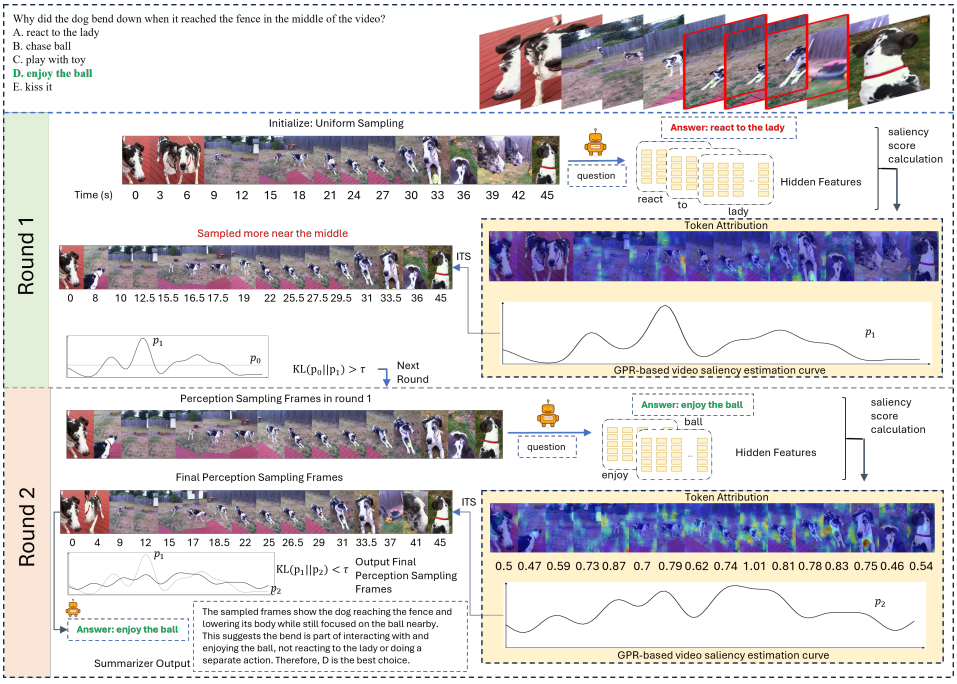


Figure 3: Qualitative example of QSTAR's progressive active perception.

Figure 3 illustrates how the iterative perception loop improves frame selection. In the first round, the initial samples miss sufficient evidence for the ball-related behavior and lead to an incorrect answer. The activation-based token attribution is then converted into frame-level saliency scores, interpolated by GPR, and used to update the sampling distribution. In the next round, the samples concentrate near the relevant middle segment of the video, allowing QSTAR to identify that the dog bends down to enjoy the ball.

3.4 Phase 3: Action Fusion

After active learning stops, QSTAR enters Phase 3 using the final k sampled frames to generate final answer. The straightforward approach is to simply feed the question and the final selection

of K sampled frames (instead of the long video) to the deployed VLM and return the output as the final answer. However, this approach may be bounded by the performance limitation of the chosen VLM. To further address those hard cases, we support optional branches and incorporate fusion along two dimensions to provide additional and yet complimentary evidence for the same question. Suppose there are $M + 1$ branch instances in total. Let branch m ($m = 0, \ldots, M$) produce sampled frames $\mathcal{F}^{(m)}$, answer $\hat{y}^{(m)}$, and optionally a short reason $R^{(m)}$, where branch 0 denotes the core saliency-based branch. When multiple branches produce different candidate answers, we perform branch-level consolidation by consensus building and conflict resolution to generate the final prediction. If different branches use different VLMs, or two branches share the same backbone VLM but use different sampling methods, we treat each such branch instance as an independent evaluator. Consider branch i is given the question, and for every output branch j, its sampled frames, its answer and reason, the branch i will evaluate each of such candidate branches and assigns a score to branch j. We perform action fusion to aggregate the results by summing the scores received by each branch. The highest-scoring branch is selected, and its answer is used as the final prediction. When all sampling approaches use the same VLM, a more efficient alternative is to provide all branch results to a single VLM judge and directly ask it to return the final answer, instead of explicitly performing consensus building and resolving conflicts across branches.

4 EXPERIMENTS

We evaluate QSTAR on five video reasoning benchmarks that cover both short and long videos, with a particular focus on long-video reasoning. For short videos with strong temporal and causal reasoning requirements, we evaluate on validation set of NeXT-QA Xiao et al. (2021) which has 4,996 QAs. We also evaluate on EgoSchema Mangalam et al. (2023), a medium-length egocentric video QA benchmark with videos around three minutes long; we use the 500-question test split with released ground-truth labels. For longer videos, we evaluate on LongVideoBench Wu et al. (2024), MLVU Zhou et al. (2024), and Video-MME Fu et al. (2025). LongVideoBench contains 1,337 QA pairs with accompanying subtitles, and the video durations range from a few minutes to roughly two hours. We use the MLVU test split, which contains about 500 videos spanning durations from several minutes to nearly two hours. Video-MME contains 2,700 QA pairs and is divided into short, medium, and long subsets with 900 QA pairs each. Unless otherwise specified, all experimental results use a fixed visual budget: 16 frames per call for NeXT-QA and EgoSchema, and 32 frames per call for LongVideoBench, MLVU, and Video-MME. For 70B+ backbones, QSTAR experiments are conducted on two H200 GPUs, and for smaller backbones, QSTAR experiments are conducted on one H200 GPU. For fair comparison, we test all open-source baselines under the same machine. Unless otherwise stated, all main-paper results use only the core saliency-based branch. Optional complementary branches and their fusion are evaluated separately in the Appendix.

Performance Comparison. We compare QSTAR with 15 open-source VLMs, 2 closed-source VLMs, and 10 long-video understanding methods in Table 1. We use InternVL3-78B as the backbone. Following a simple video-length rule, we set $K = 16$ for NeXT-QA and EgoSchema and $K = 32$ for LongVideoBench, MLVU, and Video-MME, with $\tau = 10^{-2}$ fixed throughout. These settings are chosen based on a few NeXT-QA training samples and MLVU-Dev samples to maintain a practical computational budget; Figure 4 and Table 2 report sensitivity analyses of τ and K, respectively. The CLIP branch uses $\lambda = 2.5$ and EVA-CLIP, with λ chosen from the observed CLIP similarity distributions without using evaluation labels.

With $K \ll F$, QSTAR improves InternVL3-78B by 3.3%–9.8% across all five benchmarks. It ranks first on NeXT-QA, EgoSchema, and MLVU. GPT-4o is 1.3 points higher on LongVideoBench, while Gemini 1.5 Pro is 2.8/1.3 points higher on Video-MME without/with subtitles. Among models reporting results on all five benchmarks, QSTAR achieves the highest average performance. Results with three other backbones in the Appendix show similar improvements.

Effect of frame budget K. To understand the impact of visual budget K on QSTAR performance gain, we gradually increase the frame budget up to 256 and record the best saturated performance in Table 2. We find that for the base model InternVL3-78B, its accuracy begins to saturate when the number of sampled frames reaches 64 on NeXT-QA and EgoSchema, and 128 on other benchmarks. By setting $K = 32$, QSTAR further improves the strongest backbone (InternVL3-78B) from 84.2 to 88.9 on NeXT-QA and from 79.6 to 85.4 on EgoSchema. By setting $K = 64$, QSTAR obtains further accuracy gain over InternVL3-78B (its baseline backbone VLM) from 62.3 to 66.2 on LongVideoBench, from 60.5 to 68.3 on MLVU, and on Video-MME, the saturated result reaches

Table 1: Comparison on five benchmarks. For Video-MME, results are reported in the format *w/o subtitle / w. subtitle*. The frame sampling budget $K = 16$ for NeXT-QA and EgoSchema, and $K = 32$ for LongVideoBench, MLVU, and Video-MME. The accuracy of backbone model InternVL3-78B is highlighted in blue. A method name with suffix * indicates that the results are taken from the paper or the benchmark leaderboard.

Model	NeXT-QA	EgoSchema	LongVideoBench	MLVU	Video-MME			
					Short	Medium	Long	Average
Closed-source VLMs								
GPT-4o* OpenAI (2024)	-	72.2	**66.7**	54.9	80.0/82.8	70.3/76.6	65.3/72.1	71.9/77.2
Gemini 1.5 Pro* Team et al. (2024)	-	71.1	64.0	-	81.7/84.5	**74.3/81.0**	**67.4/77.4**	**75.0/81.3**
Specialized reasoning / long-video baselines								
LongVU-7B* Shen et al. (2024)	-	37.6	-	-	-	-	-/59.5	-/60.6
VideoAgent Fan et al. (2024)	66.1	55.0	-	34.3	-	-	-	-
TraveLER Shang et al. (2024)	66.0	55.4	-	36.3	-	-	-	-
MoReVQA* Min et al. (2024)	69.2	51.7	-	-	-	-	-	-
DynFocus* Han et al. (2025)	-	-	31.8	49.6	50.9/53.9	43.7/46.0	37.7/43.6	44.1/47.8
DYTO-34B* Zhang et al. (2025c)	72.9	56.8	-	-	-	-	-	-/53.4
BOLT* Liu et al. (2025)	79.5	64.0	59.6	-	70.1/-	66.0/-	49.6/-	59.9/-
VideoRAG-72B* Luo et al. (2026)	-	-	65.4	-	81.1/-	72.9/-	73.1/-	75.7/-
FlashVID* Fan et al. (2026)	-	62.4	58.9	47.5	74.2/-	62.2/-	53.7/-	62.4/-
A.I.R.* Zou et al. (2026)	82.6	72.6	62.8	-	-	-	-	68.2/69.2
Open-source VLMs								
LLaVA-NeXT-Video-7B Zhang et al. (2024b)	39.7	39.6	43.5	26.3	36.4/40.1	30.8/38.0	27.9/17.2	31.7/31.8
VideoChat2-7B* Li et al. (2024c)	61.7	56.7	39.3	30.1	48.3/52.8	37.0/39.4	33.2/39.2	39.5/43.8
VideoLlama3-7B Zhang et al. (2025a)	67.4	52.6	52.5	43.9	80.1/80.2	63.7/69.6	54.9/61.0	66.2/70.3
LongVA-7B* Zhang et al. (2024a)	69.3	-	51.3	41.1	61.1/61.6	50.4/53.6	46.2/47.6	52.6/54.3
QwenVL-2.5-7B Bai et al. (2025)	76.8	60.7	54.7	47.7	72.1/73.6	55.7/63.9	48.2/60.9	58.7/66.1
InternVL3.5-8B Wang et al. (2025a)	76.2	62.0	53.6	47.1	69.6/74.0	56.3/63.9	47.7/61.9	57.9/66.6
LLaVA-Video-7B-Qwen2 Zhang et al. (2024c)	81.2	61.8	61.1	51.7	72.4/75.6	58.1/63.3	50.7/60.4	60.4/66.4
Aria-25.3B* Li et al. (2024b)	-	-	63.0	-	67.9/78.3	67.0/71.7	58.8/66.3	67.6/72.1
QwenVL-2.5-32B Bai et al. (2025)	76.8	64.8	53.3	48.5	73.2/77.1	61.6/68.2	54.1/68.6	63.0/71.3
LLaVA-NeXT-Video-34B Zhang et al. (2024b)	70.1	46.6	50.5	39.5	63.1/66.4	51.1/53.2	44.6/48.7	52.5/56.0
LLaVA-OneVision-72B* Li et al. (2024a)	80.2	62.0	63.2	-	76.7/79.3	62.2/66.9	66.0/62.4	66.3/69.6
InternVL3.5-38B Wang et al. (2025a)	78.9	75.4	57.0	56.4	76.1/79.1	58.6/68.9	57.4/68.3	64.0/72.1
LLaVA-Video-72B-Qwen2 Zhang et al. (2024c)	83.7	69.4	60.7	51.5	77.8/80.0	64.9/70.8	58.6/73.0	67.1/74.6
QwenVL-2.5-72B Bai et al. (2025)	79.6	70.7	56.1	53.7	75.1/77.0	66.3/73.3	59.6/70.8	67.0/73.7
InternVL3-78B Zhu et al. (2025)	84.0	76.8	56.4	55.3	78.6/80.4	65.4/73.4	56.7/67.8	66.9/73.9
QSTAR	**87.3**	**85.2**	**65.4**	**65.1**	**83.3/85.2**	**70.6/78.3**	**62.8/76.6**	**72.2/80.0**

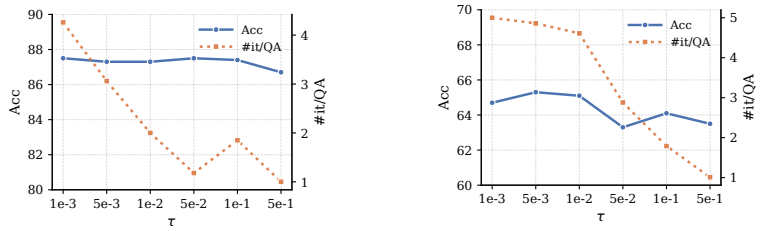
73.3/80.4 on the average score. In all five benchmarks, increasing K further yields no further gains. It is worth noting that the saturated result of QSTAR, with InternVL3-78B as its VLM backbone, surpasses the reported human accuracy on both NeXT-QA and EgoSchema.

Table 2: Frame-budget scaling results. We progressively increase the frame budget up to 256 and report the saturated accuracy on each benchmark; Table 1 uses the standard fixed-budget setting. Human results are taken from the benchmark papers Xiao et al. (2021); Mangalam et al. (2023). Video-MME results are reported as *w/o subtitle / w. subtitle*. Green numbers show the absolute improvement over the InternVL3-78B backbone.

Model	NeXT-QA	EgoSchema	LongVideoBench	MLVU	Video-MME			
					Short	Medium	Long	Average
InternVL3-78B Wang et al. (2025a)	84.2	79.6	62.3	60.5	79.6/81.2	65.7/71.9	57.0/69.4	67.4/74.2
Human Xiao et al. (2021); Mangalam et al. (2023)	88.4	76.2	-	-	-	-	-	-
QSTAR	**88.9**	**85.4**	**66.2**	**68.3**	**83.9/86.2**	**72.2/79.0**	**63.9/76.0**	**73.3/80.4**
Improvement	**+4.7**	**+5.8**	**+3.9**	**+7.8**	**+4.3/+5.0**	**+6.5/+7.1**	**+6.9/+6.6**	**+5.9/+6.2**

Sensitivity to the stopping threshold τ. Figure 4 studies how the KL-divergence threshold controls/impacts the active-learning efficiency. A smaller τ requires a smaller change between consecutive sampling distributions before stopping, so it increases the average number of iterations per question. A larger τ stops earlier and reduces computation, but may terminate before the saliency distribution is sufficiently refined. On MLVU, accuracy is best around $\tau = 5 \times 10^{-3}$ and remains close at $\tau = 10^{-2}$, while the average number of iterations decreases compared with very small thresholds. When τ is increased to 5×10^{-2} or larger, the active perception sampling stops much earlier and suffers accuracy drop, showing that long video reasoning indeed benefits from multi-round reflection and progressive refinement. On NeXT-QA, accuracy is relatively stable from 10^{-3} to 10^{-1}, indicating that shorter videos tend to require fewer refinement iterations. Therefore our default setting $\tau = 10^{-2}$ provides a strong balance between accuracy and inference cost across both short and long video benchmarks.

Inference Latency Comparison. Table 3 compares end-to-end inference latency at the performance-oriented operating point used for each method. Our goal is to evaluate whether QSTAR can attain competitive or higher reasoning accuracy while requiring substantially less inference time. For InternVL3-78B we set the number of sampled frames to be $\min\{256, 2 \times \text{video_len}\}$ where video_len



NeXT-QA MLVU

Figure 4: Ablation on the KL-divergence stopping threshold τ. Smaller τ leads to more perception rounds, while larger τ stops earlier.

Table 3: Average per-sample inference time comparison with four representative methods.

Model	NeXT-QA	EgoSchema	MLVU
GPT-4o OpenAI (2024)	23s	54s	127s
InternVL3-78B Wang et al. (2025a)	15s	50s	204s
VideoAgent Fan et al. (2024)	20s	83s	175s
TraveLER Shang et al. (2024)	101s	94s	450s
QSTAR	**7s**	**18s**	**57s**

is the video length in seconds, and for other methods, we use the same sampling rate as the one reported in their papers. We make two observations. First, QSTAR achieves a substantially more favorable accuracy–latency trade off than direct large-VLM inference and the representative agentic reasoning methods considered here: it attains comparable or higher benchmark accuracy while requiring considerably lower end-to-end inference latency. On NeXT-QA, QSTAR takes only 7s per sample, compared with 15s for InternVL3-78B and 20s for VideoAgent. The efficiency gain becomes more pronounced on longer videos: For EgoSchema and MLVU, QSTAR reduces the inference time to 18s and 57s respectively, while InternVL3-78B requires 50s and 204s. Compared with multi-step agentic methods, QSTAR is much faster with very low inference latency, compared to VideoAgent and TraveLER respectively, with $4.6\times$ and $5.2\times$ on EgoSchema and $3.07\times$ and $7.89\times$ on MLVU. These measurements show that QSTAR can reach comparable or substantially higher reasoning accuracy at markedly lower end-to-end latency, yielding a favorable accuracy-latency trade-off for long-video understanding.

Remarks. Due to the page limit, we provide in the Appendix a theoretical analysis of the QSTAR spatio-temporal active video perception sampling process with finite-round high-probability error guarantees, together with extensive ablations on the sampling budget K, stopping threshold and stopping conditions, and runtime/refinement overhead. We further provide controlled same-backbone comparisons of activation-, attention-, and CLIP-based frame scoring Sun et al. (2023), ablations of temporal interpolation methods including Linear, Quadratic, PCHIP Fritsch & Carlson (1980), and Spline Holladay (1957), analyses of the CLIP branch and fusion strategies, detailed comparison to reproduced related iterative approaches, and evaluations across different backbone VLMs, showing that in several cases, a 38B VLM powered with QSTAR can substantially outperform larger 70B+ VLM baselines.

5 Conclusion

We presented QSTAR, a training-free evidence-seeking agent framework, for improving LVU with query-adaptive frame selection with three novel contributions: *First*, QSTAR is empowered with saliency-optimized active video perception, which progressively seeks spatio-temporal visual cues as the internal perception reasoning signals to produce a saliency-refined sampling distribution. *Second*, QSTAR performs the KL-divergence based active learning to refine the sampling distribution. *Finally*, QSTAR carries out action fusion for multi-modal reasoning based on the small number of adaptively sampled frames to generate the final answer. Evaluated on five videoQA benchmarks with comparison to 10 long video specialized methods and 17 representative VLMs, we show that QSTAR consistently achieves higher performance, effectively bridging the gap between computational efficiency and spatio-temporal video perception reasoning accuracy.

REFERENCES

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

Xiangyu Bai, Bishoy Galoaa, and Sarah Ostadabbas. Hornet: Task-guided frame selection for video question answering with vision-language models. *arXiv preprint arXiv:2603.18850*, 2026.

Xiaoyi Bao, Chenwei Xie, Hao Tang, Tingyu Weng, Xiaofeng Wang, Yun Zheng, and Xingang Wang. Dynimg: Key frames with visual prompts are good representation for multi-modal video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23678–23688, 2025.

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. Himu: Hierarchical multimodal frame selection for long video question answering. *arXiv preprint arXiv:2603.18558*, 2026.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20237–20246, 2025a.

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*, pp. 19–35. Springer, 2024.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations (ICLR)*, 2025b.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. FlashVID: Efficient video large language models via training-free tree-based spatiotemporal token merging. In *The Fourteenth International Conference on Learning Representations*, 2026.

Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *arXiv preprint arXiv:2505.14321*, 2025.

Frederick N Fritsch and Ralph E Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 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*, pp. 24108–24118, 2025.

Yudong Han, Qingpei Guo, Liyuan Pan, Liu Liu, Yu Guan, and Ming Yang. Dynfocus: Dynamic cooperative network empowers llms with video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8512–8522, 2025.

John C Holladay. A smoothest curve approximation. *Mathematical tables and other aids to computation*, 11(60):233–243, 1957.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, 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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024c.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024d.

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 Computer Vision and Pattern Recognition Conference*, pp. 3318–3327, 2025.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033, 2026.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18936–18946, 2025.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 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.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ziqi Pang and Yu-Xiong Wang. Mr. video:" mapreduce" is the principle for long video understanding. *arXiv preprint arXiv:2504.16082*, 2025.

Tianyuan Qu, Longxiang Tang, Bohao Peng, Senqiao Yang, Bei Yu, and Jiaya Jia. Does your vision-language model get lost in the long video sampling dilemma? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20889–20899, 2025.

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

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 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.

Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. Traveler: A modular multi-lmm agent framework for video question-answering. *arXiv preprint arXiv:2404.01476*, 2024.

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

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26160–26169, 2025.

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*, pp. 24090–24101, 2025.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Yucheng Suo, Fan Ma, Linchao Zhu, Tianyi Wang, Fengyun Rao, and Yi Yang. From trial to triumph: Advancing long video understanding via visual context sample scaling and self-reward alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23243–23255, 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 Computer Vision and Pattern Recognition Conference*, pp. 29118–29128, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Chapter-llama: Efficient chaptering in hour-long videos with llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18947–18958, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*, 2025b.

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

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

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Weili Xu, Enxin Song, Wenhao Chai, Xuexiang Wen, Tian Ye, and Gaoang Wang. Auroralong: Bringing rnns back to efficient open-ended video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Yichang Xu, Gaowen Liu, Ramana Rao Kompella, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, and Ling Liu. A neurosymbolic agent system for compositional visual reasoning. *arXiv preprint arXiv:2506.07778*, 2025b.

Yichang Xu, Gaowen Liu, Ramana Rao Kompella, Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, and Ling Liu. A multi-agent perception-action alliance for efficient long video reasoning. *arXiv preprint arXiv:2603.14052*, 2026.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. Vca: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20168–20179, 2025.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8579–8591, 2025.

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

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

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

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025b.

Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zenghui Ding, Xianjun Yang, and Yining Sun. Beyond training: Dynamic token merging for zero-shot video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22046–22055, 2025c.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024c.

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

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

Zirui Zhu, Hailun Xu, Yang Luo, LIU YONG, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 98463–98486, 2026.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. Air: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Appendix

We organize the appendix as follows.

Section 1 provides finite-round theoretical analysis of QSTAR's reconstruction and sampling updates.

Section 2 provides detailed equation and parameters for GPR.

Section 3 describes the alternative sampling branch with CLIP similarity.

Section 4 provides additional ablation study results.

Section 5 provides the pseudocode for QSTAR active learning.

Section 6 provides failure cases and shows how we may resolve the mistakes.

Section 7 analyzes the limitations of QSTAR.

Finally, we provide AI use statement, ethics statement and reproducibility statement.

A Theoretical Analysis

Let a video contain F frames with timestamps $t_1 < \cdots < t_F$. QSTAR selects K distinct frames at each round, where $1 \le K \le F$. Here F is the total number of frames and K is the visual budget of one VLM call.

For a fixed video–question pair, let $\mathbf{s}^\star = (s_1^\star, \ldots, s_F^\star)$ be the latent query-dependent frame utilities. The update uses a fixed positive score floor $\underline{s}$:

$$0 < \underline{s} \le s_j^\star \qquad \text{for every } j \in \{1, \ldots, F\}. \tag{1}$$

Although decisive evidence may be localized, other frames can remain weakly informative through temporal and scene context. Equation (1) represents this weak dependence. The value of $\underline{s}$ may be arbitrarily small and does not imply comparable relevance across frames.

Define

$$S^\star := \sum_{j=1}^{F} s_j^\star, \qquad \pi_j^\star := \frac{s_j^\star}{S^\star}, \quad j = 1, \ldots, F. \tag{2}$$

$S^\star > 0$ is the total latent utility, and $\boldsymbol{\pi}^\star = (\pi_1^\star, \ldots, \pi_F^\star)$ is its normalized distribution.

Let $\mathbf{g} = (g_1, \ldots, g_F) \in \mathbb{R}^F$ be a vector of raw estimated utilities. The score-floor map and the distribution used by QSTAR are

$$\phi_{\underline{s}}(x) := \max\{x, \underline{s}\}, \qquad \widetilde{g}_j := \phi_{\underline{s}}(g_j), \qquad Z(\mathbf{g}) := \sum_{i=1}^{F} \widetilde{g}_i, \qquad p_j(\mathbf{g}) := \frac{\widetilde{g}_j}{Z(\mathbf{g})}. \tag{3}$$

Since $Z(\mathbf{g}) \ge F\underline{s} > 0$, the update is defined for every $\mathbf{g}$. No probability mass is mixed in after normalization. For probability vectors $\mathbf{p}$ and $\mathbf{q}$, define

$$\|\mathbf{p} - \mathbf{q}\|_1 := \sum_{j=1}^{F} |p_j - q_j|, \qquad \mathrm{TV}(\mathbf{p}, \mathbf{q}) := \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1, \tag{4}$$

and

$$D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) := \sum_{j=1}^{F} p_j \log \frac{p_j}{q_j}, \tag{5}$$

where $0 \log(0/q_j)$ is interpreted as zero.

Theorem A.1 (Normalization under a positive score floor)**.** *For any* $\mathbf{g} \in \mathbb{R}^F$*, let*

$$E(\mathbf{g}) := \sum_{j=1}^{F} \left| \phi_{\underline{s}}(g_j) - s_j^\star \right|. \tag{6}$$

Then

$$|Z(\mathbf{g}) - S^\star| \le E(\mathbf{g}), \tag{7}$$

$$\|\mathbf{p}(\mathbf{g}) - \boldsymbol{\pi}^\star\|_1 \le \frac{2E(\mathbf{g})}{S^\star}, \tag{8}$$

$$\mathrm{TV}(\mathbf{p}(\mathbf{g}), \boldsymbol{\pi}^\star) \le \frac{E(\mathbf{g})}{S^\star}. \tag{9}$$

If $\|\mathbf{g} - \mathbf{s}^\star\|_\infty \le a$ *for some* $a \ge 0$*, then*

$$E(\mathbf{g}) \le Fa. \tag{10}$$

Every component of the updated distribution is positive, with

$$\min_{1\le j\le F} p_j(\mathbf{g}) \ge \frac{\underline{s}}{Z(\mathbf{g})} > 0. \tag{11}$$

Proof. Write $\widetilde{\mathbf{g}} = (\widetilde{g}_1, \dots, \widetilde{g}_F)$ and $Z = Z(\mathbf{g})$. By the triangle inequality,

$$\begin{aligned} |Z - S^\star| &= \left| \sum_{j=1}^{F} (\widetilde{g}_j - s_j^\star) \right| \\ &\le \sum_{j=1}^{F} |\widetilde{g}_j - s_j^\star| = E(\mathbf{g}), \end{aligned}$$

which proves (7).

For the normalized vectors, add and subtract $\widetilde{\mathbf{g}}/S^\star$ to obtain

$$\begin{aligned} \left\| \frac{\widetilde{\mathbf{g}}}{Z} - \frac{\mathbf{s}^\star}{S^\star} \right\|_1 &\le \left\| \frac{\widetilde{\mathbf{g}}}{Z} - \frac{\widetilde{\mathbf{g}}}{S^\star} \right\|_1 + \left\| \frac{\widetilde{\mathbf{g}}}{S^\star} - \frac{\mathbf{s}^\star}{S^\star} \right\|_1 \\ &= \frac{|S^\star - Z|}{S^\star} + \frac{E(\mathbf{g})}{S^\star} \\ &\le \frac{2E(\mathbf{g})}{S^\star}. \end{aligned}$$

This proves (8). Equation (9) follows from the definition of total variation.

$\phi_{\underline{s}}$ is the Euclidean projection onto $[\underline{s}, \infty)$. Since $s_j^\star \in [\underline{s}, \infty)$ by (1), projection cannot increase the distance to $s_j^\star$:

$$\left|\phi_{\underline{s}}(g_j) - s_j^\star\right| \le |g_j - s_j^\star|. \tag{12}$$

$$E(\mathbf{g}) \le \sum_{j=1}^{F} |g_j - s_j^\star| \le F\|\mathbf{g} - \mathbf{s}^\star\|_\infty \le Fa,$$

which proves (10). Since $\widetilde{g}_j \ge \underline{s}$ and $Z > 0$, $p_j(\mathbf{g}) = \widetilde{g}_j/Z \ge \underline{s}/Z$, proving (11). □

Observation model and GPR estimator. Rounds are indexed by $r \in \mathbb{N}$. At round r, let $J_r \subseteq \{1, \dots, F\}$ be the set of K frames selected from the history preceding that round; J_r may be adaptive. For every $j \in J_r$, the activation-based score is modeled as

$$Y_{j,r} = s_j^\star + b_{j,r} + \xi_{j,r}. \tag{13}$$

The term $b_{j,r}$ is the systematic context-dependent deviation produced by the other selected frames and by the answer at round r. The residual $\xi_{j,r}$ is conditionally mean zero and σ-sub-Gaussian: for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda \xi_{j,r}) \mid \mathcal{F}_{j,r}] \le \exp\left(\frac{\lambda^2 \sigma^2}{2}\right). \tag{14}$$

Here $\sigma > 0$ is fixed, and $\mathcal{F}_{j,r}$ contains the history available immediately before $\xi_{j,r}$ is observed. The conditional form permits adaptive selection and dependence across rounds.

Let

$$X_r := \bigcup_{q=1}^{r} J_q, \qquad N_j(r) := \sum_{q=1}^{r} \mathbf{1}\{j \in J_q\}, \quad j \in X_r, \tag{15}$$

where X_r is the set of frames observed by round r, $N_j(r)$ is the number of observations of frame j, and $\mathbf{1}\{\cdot\}$ is the indicator function. The cache stores the running mean

$$\overline{Y}_{j,r} := \frac{1}{N_j(r)} \sum_{q \le r : j \in J_q} Y_{j,q}. \tag{16}$$

The corresponding averaged bias, averaged residual, and maximum cached bias are

$$\bar{b}_{j,r} := \frac{1}{N_j(r)} \sum_{q \le r : j \in J_q} b_{j,q}, \qquad \bar{\xi}_{j,r} := \frac{1}{N_j(r)} \sum_{q \le r : j \in J_q} \xi_{j,q}, \qquad B_r := \max_{j \in X_r} |\bar{b}_{j,r}|. \tag{17}$$

Let Θ be a finite collection of RBF-kernel hyperparameters. For each $\theta \in \Theta$, set

$$k_\theta(t, t') := \alpha_\theta^2 \exp\left(-\frac{(t-t')^2}{2\ell_\theta^2}\right), \tag{18}$$

where $\alpha_\theta > 0$ and $\ell_\theta > 0$ are the kernel amplitude and length scale. Let $\mathcal{H}_\theta$ be the RKHS of k_θ. For each θ, assume that there is an $f_\theta \in \mathcal{H}_\theta$ satisfying

$$\|f_\theta\|_{\mathcal{H}_\theta} \le R_\theta, \qquad \max_{1 \le j \le F} |s_j^\star - f_\theta(t_j)| \le \zeta_\theta, \tag{19}$$

where $R_\theta \ge 0$ bounds the RKHS norm and $\zeta_\theta \ge 0$ is the approximation error on the finite temporal grid.

Write $X_r = \{i_1, \dots, i_{n_r}\}$, where $n_r := |X_r|$, and define

$$\begin{aligned} \mathbf{K}_{\theta,r} &:= \big(k_\theta(t_{i_a}, t_{i_b})\big)_{a,b=1}^{n_r}, \\ \mathbf{k}_{\theta,r}(t) &:= \big(k_\theta(t_{i_1}, t), \dots, k_\theta(t_{i_{n_r}}, t)\big)^\top, \\ \overline{\mathbf{Y}}_r &:= \big(\overline{Y}_{i_1,r}, \dots, \overline{Y}_{i_{n_r},r}\big)^\top. \end{aligned} \tag{20}$$

For a nugget $\nu_{\theta,r} > 0$, the GPR weight vector and prediction are

$$\mathbf{w}_{\theta,r}(t) := \big(\mathbf{K}_{\theta,r} + \nu_{\theta,r}\mathbf{I}\big)^{-1} \mathbf{k}_{\theta,r}(t), \tag{21}$$

$$g_{\theta,r}(j) := \mathbf{w}_{\theta,r}(t_j)^\top \overline{\mathbf{Y}}_r, \qquad j = 1, \dots, F. \tag{22}$$

The identity matrix $\mathbf{I}$ has dimension $n_r \times n_r$. The hyperparameter $\widehat{\theta}_r \in \Theta$ may depend on all observations available at round r. Set

$$\mathbf{g}_r := \big(g_{\widehat{\theta}_r,r}(1), \dots, g_{\widehat{\theta}_r,r}(F)\big), \qquad \mathbf{p}_r := \mathbf{p}(\mathbf{g}_r). \tag{23}$$

Define

$$\Lambda_{\theta,r} := \max_{1 \le j \le F} \|\mathbf{w}_{\theta,r}(t_j)\|_1, \tag{24}$$

$$P_{\theta,r} := \max_{1 \le j \le F} \Big[k_\theta(t_j, t_j) - \mathbf{k}_{\theta,r}(t_j)^\top \big(\mathbf{K}_{\theta,r} + \nu_{\theta,r}\mathbf{I}\big)^{-1} \mathbf{k}_{\theta,r}(t_j)\Big]^{1/2}, \tag{25}$$

$$h_r := \max_{1 \le j \le F} \min_{i \in X_r} |t_j - t_i|. \tag{26}$$

$\Lambda_{\theta,r}$ is the maximum ℓ_1 norm of the GPR weights, $P_{\theta,r}$ is the maximum regularized power function on the frame grid, and h_r is the maximum distance from a frame to the nearest observed frame.

Theorem A.2 (Finite-round reconstruction and distribution stability)**.** *Fix* $0 < \delta < 1$. *For* $n \ge 1$, *define the anytime radius*

$$e_n(\delta) := \sigma \sqrt{\frac{2}{n} \log\left(\frac{\pi^2 F n^2}{3\delta}\right)}, \qquad \varepsilon_r(\delta) := \max_{j \in X_r} e_{N_j(r)}(\delta). \tag{27}$$

Let $\theta_r := \widehat{\theta}_r$ and set

$$a_r := R_{\theta_r} P_{\theta_r,r} + (1 + \Lambda_{\theta_r,r})\zeta_{\theta_r} + \Lambda_{\theta_r,r}\big(B_r + \varepsilon_r(\delta)\big). \tag{28}$$

There is an event $\mathcal{E}_\delta$ with $\Pr(\mathcal{E}_\delta) \geq 1 - \delta$ on which the following bounds hold simultaneously for all $r \geq 1$:

$$\|\mathbf{g}_r - \mathbf{s}^\star\|_\infty \leq a_r, \tag{29}$$

$$\mathrm{TV}(\mathbf{p}_r, \boldsymbol{\pi}^\star) \leq \frac{F a_r}{S^\star}. \tag{30}$$

For $r \geq 2$, let

$$\beta_{r-1} := \min_{1 \leq j \leq F} p_{r-1,j}. \tag{31}$$

The score floor gives $\beta_{r-1} \geq \underline{s}/Z(\mathbf{g}_{r-1}) > 0$, and on the same event

$$D_{\mathrm{KL}}(\mathbf{p}_r \| \mathbf{p}_{r-1}) \leq \frac{4F^2}{\beta_{r-1}(S^\star)^2}(a_r + a_{r-1})^2 \leq \frac{4F^2 Z(\mathbf{g}_{r-1})}{\underline{s}(S^\star)^2}(a_r + a_{r-1})^2. \tag{32}$$

For every $\theta \in \Theta$ and finite round r,

$$P_{\theta,r} \leq \min\left\{\alpha_\theta, \sqrt{\nu_{\theta,r} + \frac{\alpha_\theta^2 h_r^2}{\ell_\theta^2}}\right\}. \tag{33}$$

Proof. Cached residuals. Fix a frame j and list its residuals in observation order as $\xi'_{j,1}, \xi'_{j,2}, \ldots$. Since the decision to observe frame j is made before the corresponding residual is revealed, (14) applies to this adaptively selected sequence. The exponential supermartingale bound gives, for every $n \geq 1$ and every $x > 0$,

$$\Pr\left(\left|\frac{1}{n}\sum_{q=1}^{n} \xi'_{j,q}\right| > x\right) \leq 2\exp\left(-\frac{nx^2}{2\sigma^2}\right). \tag{34}$$

If j is observed fewer than n times, the associated event is empty. Substituting $x = e_n(\delta)$ into (34) yields

$$\Pr\left(\left|\frac{1}{n}\sum_{q=1}^{n} \xi'_{j,q}\right| > e_n(\delta)\right) \leq \frac{6\delta}{\pi^2 F n^2}.$$

A union bound over all F frames and all $n \geq 1$, together with $\sum_{n=1}^{\infty} n^{-2} = \pi^2/6$, shows that an event $\mathcal{E}_\delta$ with $\Pr(\mathcal{E}_\delta) \geq 1 - \delta$ exists on which

$$|\bar{\xi}_{j,r}| \leq e_{N_j(r)}(\delta) \leq \varepsilon_r(\delta) \quad \text{for every } r \text{ and every } j \in X_r. \tag{35}$$

The bound is uniform over finite rounds and therefore also holds at a data-dependent stopping round.

GPR reconstruction error. Fix r, θ, and j, and write

$$\mathbf{w} := \mathbf{w}_{\theta,r}(t_j), \qquad k_j := k_\theta(\cdot, t_j), \qquad k_{i_a} := k_\theta(\cdot, t_{i_a}).$$

By the reproducing property and the Cauchy–Schwarz inequality,

$$\left|f_\theta(t_j) - \sum_{a=1}^{n_r} w_a f_\theta(t_{i_a})\right| \leq R_\theta \left\|k_j - \sum_{a=1}^{n_r} w_a k_{i_a}\right\|_{\mathcal{H}_\theta}. \tag{36}$$

For $\mathbf{c} \in \mathbb{R}^{n_r}$, define

$$\mathcal{L}_{\theta,j,r}(\mathbf{c}) := \left\|k_j - \sum_{a=1}^{n_r} c_a k_{i_a}\right\|_{\mathcal{H}_\theta}^2 + \nu_{\theta,r}\|\mathbf{c}\|_2^2. \tag{37}$$

Expanding the squared norm gives

$$\mathcal{L}_{\theta,j,r}(\mathbf{c}) = k_\theta(t_j, t_j) - 2\mathbf{c}^\top \mathbf{k}_{\theta,r}(t_j) + \mathbf{c}^\top(\mathbf{K}_{\theta,r} + \nu_{\theta,r}\mathbf{I})\mathbf{c}.$$

The matrix $\mathbf{K}_{\theta,r} + \nu_{\theta,r}\mathbf{I}$ is positive definite, with unique minimizer $\mathbf{c} = \mathbf{w}_{\theta,r}(t_j)$, and the minimum equals

$$k_\theta(t_j, t_j) - \mathbf{k}_{\theta,r}(t_j)^\top (\mathbf{K}_{\theta,r} + \nu_{\theta,r}\mathbf{I})^{-1}\mathbf{k}_{\theta,r}(t_j). \tag{38}$$

Since (38) includes the nonnegative term $\nu_{\theta,r}\|\mathbf{w}\|_2^2$, its square root bounds the RKHS norm in (36) and is at most $P_{\theta,r}$.

Let $d_{\theta,j} := s_j^\star - f_\theta(t_j)$. By (19), $|d_{\theta,j}| \le \zeta_\theta$ for every j. From (13) and (17),

$$\overline{Y}_{i_a,r} = f_\theta(t_{i_a}) + d_{\theta,i_a} + \overline{b}_{i_a,r} + \overline{\xi}_{i_a,r}. \tag{39}$$

On $\mathcal{E}_\delta$,

$$\begin{aligned}
|g_{\theta,r}(j) - s_j^\star| &\le \left|\sum_{a=1}^{n_r} w_a f_\theta(t_{i_a}) - f_\theta(t_j)\right| \\
&\quad + \left|\sum_{a=1}^{n_r} w_a d_{\theta,i_a} - d_{\theta,j}\right| \\
&\quad + \left|\sum_{a=1}^{n_r} w_a(\overline{b}_{i_a,r} + \overline{\xi}_{i_a,r})\right| \\
&\le R_\theta P_{\theta,r} + (1 + \|\mathbf{w}\|_1)\zeta_\theta + \|\mathbf{w}\|_1\big(B_r + \varepsilon_r(\delta)\big) \\
&\le R_\theta P_{\theta,r} + (1 + \Lambda_{\theta,r})\zeta_\theta + \Lambda_{\theta,r}\big(B_r + \varepsilon_r(\delta)\big).
\end{aligned}$$

The third term contains the required weight factor $\|\mathbf{w}\|_1$. Since $\mathcal{E}_\delta$ does not depend on θ, the inequality also holds for the data-dependent choice $\theta_r = \widehat{\theta}_r$. Maximizing over j proves (29). Theorem A.1 and $E(\mathbf{g}_r) \le F a_r$ then give (30).

Power-function bound. Evaluating $\mathcal{L}_{\theta,j,r}$ at $\mathbf{c} = \mathbf{0}$ gives $P_{\theta,r} \le \alpha_\theta$. For each j, choose $i \in X_r$ satisfying $|t_j - t_i| \le h_r$ and evaluate $\mathcal{L}_{\theta,j,r}$ at the vector that assigns coefficient one to i and zero to every other observed frame. For the RBF kernel,

$$\begin{aligned}
\|k_\theta(\,\cdot\,, t_j) - k_\theta(\,\cdot\,, t_i)\|_{\mathcal{H}_\theta}^2 &= 2\alpha_\theta^2\left[1 - \exp\left(-\frac{(t_j - t_i)^2}{2\ell_\theta^2}\right)\right] \\
&\le \frac{\alpha_\theta^2}{\ell_\theta^2}(t_j - t_i)^2,
\end{aligned}$$

where $1 - e^{-x} \le x$ for $x \ge 0$. The variational characterization gives

$$P_{\theta,r}^2 \le \nu_{\theta,r} + \frac{\alpha_\theta^2 h_r^2}{\ell_\theta^2},$$

which proves (33).

Successive KL divergence. For probability vectors $\mathbf{p}$ and $\mathbf{q}$ with $q_j > 0$ for all j, the inequality $\log x \le x - 1$ implies

$$D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) \le \sum_{j=1}^{F} \frac{(p_j - q_j)^2}{q_j} \le \frac{1}{\min_j q_j}\|\mathbf{p} - \mathbf{q}\|_1^2. \tag{40}$$

The score floor gives $\beta_{r-1} > 0$. Also,

$$\begin{aligned}
\|\mathbf{p}_r - \mathbf{p}_{r-1}\|_1 &\le \|\mathbf{p}_r - \boldsymbol{\pi}^\star\|_1 + \|\mathbf{p}_{r-1} - \boldsymbol{\pi}^\star\|_1 \\
&\le \frac{2F}{S^\star}(a_r + a_{r-1}).
\end{aligned}$$

Applying (40) with $\mathbf{p} = \mathbf{p}_r$, $\mathbf{q} = \mathbf{p}_{r-1}$, and $\min_j q_j = \beta_{r-1}$ proves the first inequality in (32). The second follows from $\beta_{r-1} \ge \underline{s}/Z(\mathbf{g}_{r-1})$. □

Remark on the stopping rule. The implication in (32) is one-sided. Small values of a_{r-1} and a_r imply a small successive KL divergence, but a small KL divergence need not imply accurate reconstruction. For example, if $\mathbf{g}_{r-1} = \mathbf{g}_r = c\mathbf{1}$ for some $c > \underline{s}$, then $D_{\mathrm{KL}}(\mathbf{p}_r \| \mathbf{p}_{r-1}) = 0$. The common distribution is uniform and may be far from a nonuniform $\boldsymbol{\pi}^\star$. The stopping rule measures stability between successive updates, not error relative to $\boldsymbol{\pi}^\star$.

The score floor makes the KL denominator positive but does not force a deterministic quantile rule to select every frame. Theorem A.2 is finite-round and does not use an infinite-visitation assumption.

Deterministic inverse-transform selection. For a probability vector $\mathbf{p} = (p_1, \ldots, p_F)$, define its cumulative distribution by

$$C_{\mathbf{p}}(0) := 0, \qquad C_{\mathbf{p}}(j) := \sum_{i=1}^{j} p_i, \quad j = 1, \ldots, F. \tag{41}$$

QSTAR uses the fixed quantile levels

$$z_\ell := \frac{\ell - \frac{1}{2}}{K}, \qquad \ell = 1, \ldots, K, \tag{42}$$

and maps each level to the first frame whose cumulative probability reaches that level:

$$a_\ell(\mathbf{p}) := \min\{j \in \{1, \ldots, F\} : C_{\mathbf{p}}(j) \geq z_\ell\}. \tag{43}$$

Let $A_K(\mathbf{p})$ denote the output after the uniqueness correction. This correction retains the distinct indices in (43) and fills only the repeated positions. If $a_1(\mathbf{p}), \ldots, a_K(\mathbf{p})$ are distinct,

$$A_K(\mathbf{p}) = \{a_1(\mathbf{p}), \ldots, a_K(\mathbf{p})\}. \tag{44}$$

Total variation bounds the change in every prefix probability and therefore controls the indices returned by inverse-transform selection. We use the selection under $\boldsymbol{\pi}^\star$ as the reference and define

$$a_\ell^\star := a_\ell(\boldsymbol{\pi}^\star), \qquad \ell = 1, \ldots, K, \tag{45}$$

Assume that these K indices are distinct. Their minimum distance from the corresponding CDF boundaries is

$$\mu_{\mathrm{sel}} := \min_{1 \leq \ell \leq K} \left\{ z_\ell - C_{\boldsymbol{\pi}^\star}(a_\ell^\star - 1), C_{\boldsymbol{\pi}^\star}(a_\ell^\star) - z_\ell \right\}, \tag{46}$$

and we assume $\mu_{\mathrm{sel}} > 0$.

For the evidence-hit statement, let $\mathcal{I} \subseteq \{1, \ldots, F\}$ contain a contiguous block

$$\mathcal{B} := \{L, L+1, \ldots, U\} \subseteq \mathcal{I}, \qquad 1 \leq L \leq U \leq F. \tag{47}$$

Let

$$M_{\mathcal{B}} := \sum_{j=L}^{U} \pi_j^\star = C_{\boldsymbol{\pi}^\star}(U) - C_{\boldsymbol{\pi}^\star}(L-1) \tag{48}$$

be the mass of this block under $\boldsymbol{\pi}^\star$. Set

$$\mu_{\mathcal{B}} := \frac{1}{2}\left(M_{\mathcal{B}} - \frac{1}{K}\right). \tag{49}$$

The evidence-hit result below assumes $M_{\mathcal{B}} > 1/K$, or equivalently $\mu_{\mathcal{B}} > 0$. For a fixed round r, define the events

$$\mathcal{G}_r := \left\{ \frac{F a_r}{S^\star} < \mu_{\mathrm{sel}} \right\}, \tag{50}$$

$$\mathcal{H}_r := \left\{ \frac{F a_r}{S^\star} < \mu_{\mathcal{B}} \right\}. \tag{51}$$

Theorem A.3 (Evidence-hit guarantees)**.** *On $\mathcal{E}_\delta$ from Theorem A.2,*

$$\mathcal{G}_r \quad \Longrightarrow \quad A_K(\mathbf{p}_r) = A_K(\boldsymbol{\pi}^\star). \tag{52}$$

If $M_{\mathcal{B}} > 1/K$, then, on the same event,

$$\mathcal{H}_r \quad \Longrightarrow \quad A_K(\mathbf{p}_r) \cap \mathcal{I} \neq \varnothing. \tag{53}$$

Under the same condition, for every fixed r,

$$\Pr(A_K(\mathbf{p}_r) \cap \mathcal{I} \neq \varnothing) \geq 1 - \delta - \Pr(\mathcal{H}_r^c). \tag{54}$$

If $F a_r / S^\star < \mu_{\mathcal{B}}$ holds surely at round r, then $\Pr(\mathcal{H}_r^c) = 0$ and the lower bound is $1 - \delta$.

Proof. Selection recovery. On $\mathcal{E}_\delta \cap \mathcal{G}_r$, Theorem A.2 gives

$$\mathrm{TV}(\mathbf{p}_r, \boldsymbol{\pi}^\star) < \mu_{\mathrm{sel}}. \tag{55}$$

Each prefix $\{1, \dots, j\}$ is an event on the finite frame space. The variational characterization of total variation therefore yields

$$|C_{\mathbf{p}_r}(j) - C_{\boldsymbol{\pi}^\star}(j)| \le \mathrm{TV}(\mathbf{p}_r, \boldsymbol{\pi}^\star) < \mu_{\mathrm{sel}} \tag{56}$$

for every $j \in \{0, \dots, F\}$.

Fix $\ell \in \{1, \dots, K\}$ and let $a = a^\star_\ell$. From the definition of μ_{sel} and (56),

$$C_{\mathbf{p}_r}(a-1) < C_{\boldsymbol{\pi}^\star}(a-1) + \mu_{\mathrm{sel}} \le z_\ell,$$
$$C_{\mathbf{p}_r}(a) > C_{\boldsymbol{\pi}^\star}(a) - \mu_{\mathrm{sel}} \ge z_\ell.$$

The first index at which $C_{\mathbf{p}_r}$ reaches z_ℓ is therefore a, so $a_\ell(\mathbf{p}_r) = a^\star_\ell$. Applying the same argument to each ℓ proves (52). The recovered indices are distinct, and the uniqueness correction leaves them unchanged.

Evidence hit. Put

$$x := C_{\boldsymbol{\pi}^\star}(L-1), \qquad y := C_{\boldsymbol{\pi}^\star}(U), \qquad m := \frac{x+y}{2}.$$

By (48), $y - x = M_{\mathcal{B}}$. The fixed levels $z_1, \dots, z_K$ are the midpoints of the K intervals $[0, 1/K], \dots, [(K-1)/K, 1]$. Some z_{ℓ_0} therefore satisfies $|z_{\ell_0} - m| \le 1/(2K)$. It follows that

$$z_{\ell_0} - x \ge \frac{M_{\mathcal{B}}}{2} - \frac{1}{2K} = \mu_{\mathcal{B}},$$
$$y - z_{\ell_0} \ge \frac{M_{\mathcal{B}}}{2} - \frac{1}{2K} = \mu_{\mathcal{B}}. \tag{57}$$

On $\mathcal{E}_\delta \cap \mathcal{H}_r$, the distribution bound in Theorem A.2 gives $\mathrm{TV}(\mathbf{p}_r, \boldsymbol{\pi}^\star) < \mu_{\mathcal{B}}$. Applying the variational characterization of total variation to the prefixes $\{1, \dots, L-1\}$ and $\{1, \dots, U\}$ gives

$$C_{\mathbf{p}_r}(L-1) < x + \mu_{\mathcal{B}} \le z_{\ell_0} \quad \text{and} \quad C_{\mathbf{p}_r}(U) > y - \mu_{\mathcal{B}} \ge z_{\ell_0}.$$

The first index at which $C_{\mathbf{p}_r}$ reaches z_{ℓ_0} lies between L and U, and $a_{\ell_0}(\mathbf{p}_r) \in \mathcal{B} \subseteq \mathcal{I}$. The uniqueness correction retains this index, and $A_K(\mathbf{p}_r) \cap \mathcal{I} \neq \varnothing$ on $\mathcal{E}_\delta \cap \mathcal{H}_r$.

The union bound gives

$$\Pr(\mathcal{E}_\delta \cap \mathcal{H}_r) \ge 1 - \Pr(\mathcal{E}_\delta^{\mathrm{c}}) - \Pr(\mathcal{H}_r^{\mathrm{c}}) \ge 1 - \delta - \Pr(\mathcal{H}_r^{\mathrm{c}}),$$

which proves (54). □

Theorem A.4 (Lower additional cost of activation-based saliency scoring)**.** *Let K be the number of sampled frames, P be the number of visual tokens per frame, and $n_v = KP$ be the total number of visual tokens. Let n_p and n_a denote the number of prompt tokens and generated answer tokens, and let $n = n_v + n_p + n_a$. Let d be the hidden dimension, L be the number of transformer layers, and H be the number of attention heads. Suppose the VLM forward pass has already produced the final-layer hidden states used by QSTAR.*

A dense attention-map-based saliency method that extracts and aggregates token-token attention maps over all layers and heads has additional time and memory complexity

$$T_{\mathrm{att}} = \Omega(LHn^2), \qquad S_{\mathrm{att}} = \Omega(LHn^2).$$

In contrast, the proposed activation-based saliency score has additional complexity

$$T_{\mathrm{act}} = O(n_v d|\mathcal{B}|), \qquad S_{\mathrm{act}} = O(n_v d + d|\mathcal{B}| + n_v|\mathcal{B}|),$$

where $\mathcal{B}$ is the set of unique LM-head directions used for the selected answer and context tokens. Therefore, for a fixed backbone and a growing visual sequence length, activation-based scoring is linear in n_v, while dense attention-map scoring is quadratic in the full token length n.

Proof. A dense attention-map-based saliency method needs to access attention matrices

$$A^{(\ell,h)} \in \mathbb{R}^{n\times n}, \qquad \ell = 1,\dots,L, \quad h = 1,\dots,H.$$

Thus, extracting and aggregating these maps over all layers and heads requires at least $\Omega(LHn^2)$ time and $\Omega(LHn^2)$ memory.

For activation-based saliency scoring, let

$$H_v \in \mathbb{R}^{n_v\times d}$$

be the final-layer visual-token hidden states, and let

$$W_{\mathcal{B}} \in \mathbb{R}^{|\mathcal{B}|\times d}$$

contain the LM-head directions used by the scorer. The required activation maps are obtained by

$$H_v W_{\mathcal{B}}^{\top} \in \mathbb{R}^{n_v\times|\mathcal{B}|}.$$

This matrix product costs $O(n_v d|\mathcal{B}|)$ time. Storing the visual hidden states, selected LM-head directions, and projected activation scores costs

$$O(n_v d + d|\mathcal{B}| + n_v|\mathcal{B}|).$$

Thus, after the same VLM forward pass, activation-based scoring avoids dense token-token attention-map extraction and has lower additional asymptotic cost for long visual sequences. □

Proposition A.5 (Activation scores encode answer-conditioned LM-head alignment)**.** *Let a_m be a selected generated answer token, with LM-head direction $w(a_m) \in \mathbb{R}^d$, and let $h_j \in \mathbb{R}^d$ be the final-layer hidden state of visual token j. The LM-head alignment between visual token j and answer token a_m is*

$$\ell_{m,j} = h_j^{\top} w(a_m).$$

Therefore, the activation score

$$r_{m,j} = \mathrm{ReLU}\big(h_j^{\top} w(a_m)\big)$$

measures their positive LM-head alignment. Because the direction $w(a_m)$ depends on the generated answer token, this score is answer-conditioned. Attention weights alone do not determine this alignment.

Proof. The LM head maps a hidden state h to the logit of token y by

$$z_y(h) = w(y)^{\top} h.$$

Applying this readout to the visual-token state h_j for answer token a_m gives

$$z_{a_m}(h_j) = h_j^{\top} w(a_m).$$

Taking its positive part gives $r_{m,j}$. A large value means that the visual-token representation is strongly aligned with the LM-head direction of the generated answer token.

Attention weights describe how value representations are mixed, but not what those representations contain. For example, an attention output has the form

$$o_i = \sum_k \alpha_{ik} v_k.$$

Even with fixed attention weights α_{ik}, the quantity $w(a_m)^{\top} v_j$ can be positive or negative depending on v_j. Therefore, attention weights alone cannot determine whether the content at a visual-token position aligns positively with the generated answer. □

Remark (scope and empirical validation of Proposition A.5). Proposition A.5 characterizes an *answer-conditioned LM-head alignment*: $h_j^\top w(a_m)$ measures the compatibility between the final-layer representation of visual token j and the vocabulary direction of answer token a_m. We use this quantity as a frame-ranking signal rather than as a causal attribution measure. Empirically, on the NeXT-QA examples with human temporal evidence annotations from NeXT-GQA, frames inside the annotated evidence window receive higher normalized activation scores in both correctly and incorrectly answered examples ($+0.27$ and $+0.24$ standard deviations, respectively). This indicates that the signal remains informative for prioritizing evidence-bearing frames even when the current answer is imperfect. Intuitively, an incorrect answer can still be produced from visual features, entities, or actions that occur near the relevant event, so answer-conditioned alignment can retain useful temporal information without requiring the provisional answer itself to be correct. Consistent with this interpretation, the activation score achieves a ranking AP of 0.339 and improves substantially over random selection in mIoP (27.9 vs. 20.5) and IoP@0.5 (28.0 vs. 19.1). These results support its use as an empirical sampling signal for identifying frames that are informative for subsequent reasoning.

Theorem A.6 (Least-squares removal of context-aligned activation)**.** *For a selected answer token a_m, let*

$$r_m = \mathrm{ReLU}(H_v w(a_m)) \in \mathbb{R}^{n_v}$$

be the raw visual activation map. Let $u_m \in \mathbb{R}^{n_v}$ be the context-induced visual map computed from prompt tokens and previously generated answer tokens. Define

$$\beta_m = \arg\min_{\beta \in \mathbb{R}} \|r_m - \beta u_m\|_2^2, \qquad \tilde{r}_m = [r_m - \max(\beta_m, 0) u_m]_+.$$

If $u_m \neq 0$ and $\beta_m \geq 0$, then the residual

$$e_m = r_m - \beta_m u_m$$

satisfies

$$e_m^\top u_m = 0.$$

Thus, the residualization step removes the least-squares component of the raw activation map aligned with the constructed context map before nonnegative clipping.

Proof. The scalar least-squares problem

$$\min_{\beta \in \mathbb{R}} \|r_m - \beta u_m\|_2^2$$

has the closed-form solution

$$\beta_m = \frac{r_m^\top u_m}{u_m^\top u_m}$$

when $u_m \neq 0$. If $\beta_m \geq 0$, then

$$e_m = r_m - \beta_m u_m.$$

Taking the inner product with u_m gives

$$e_m^\top u_m = r_m^\top u_m - \frac{r_m^\top u_m}{u_m^\top u_m} u_m^\top u_m = 0.$$

Therefore, the residual is orthogonal to the context-induced map and removes the linear component of the raw activation map that can be explained by context. If $\beta_m < 0$, QSTAR uses $\max(\beta_m, 0) = 0$, which prevents subtraction in the opposite direction. The final clipping $[\cdot]_+$ keeps only positive answer-supporting evidence. □

These results characterize the efficiency and answer-conditioned structure of the proposed scoring procedure.

B GAUSSIAN-PROCESS ESTIMATION OF FRAME SCORES

Let a video contain F frames. Frame $f \in \{1, \dots, F\}$ is assigned the normalized temporal coordinate

$$t_f := \begin{cases} (f-1)/(F-1), & F > 1, \\ 0, & F = 1. \end{cases} \tag{58}$$

Thus, the regressor uses normalized frame indices rather than timestamps in seconds. At refinement round r, suppose scores have been cached for the frames

$$\mathcal{O}_r = \{i_1, \dots, i_{n_r}\} \subseteq \{1, \dots, F\}, \tag{59}$$

where $n_r := |\mathcal{O}_r|$. Let $y_{a,r}$ denote the cached score of frame i_a.

The cached scores are standardized before fitting. Define

$$\overline{y}_r := \frac{1}{n_r} \sum_{a=1}^{n_r} y_{a,r}, \qquad s_r := \left[\frac{1}{n_r} \sum_{a=1}^{n_r} (y_{a,r} - \overline{y}_r)^2 \right]^{1/2}. \tag{60}$$

When $n_r > 1$ and $s_r \geq 10^{-6}$, set

$$z_{a,r} := \frac{y_{a,r} - \overline{y}_r}{s_r}, \qquad a = 1, \dots, n_r. \tag{61}$$

The Gaussian process has zero mean and the unit-amplitude radial basis function kernel

$$k_\ell(t, t') := \exp\left[-\frac{(t-t')^2}{2\ell^2} \right], \tag{62}$$

where $\ell > 0$ is the temporal length scale. For an observation-noise standard deviation $\sigma_n > 0$, set

$$\begin{aligned} \mathbf{K}_{\ell,r} &:= \left(k_\ell(t_{i_a}, t_{i_b}) \right)_{a,b=1}^{n_r}, \\ \mathbf{A}_{\ell,\sigma_n,r} &:= \mathbf{K}_{\ell,r} + (\sigma_n^2 + 10^{-8}) \mathbf{I}_{n_r}, \end{aligned} \tag{63}$$

where $\mathbf{I}_{n_r}$ is the $n_r \times n_r$ identity matrix and 10^{-8} is a numerical jitter. With $\mathbf{z}_r := (z_{1,r}, \dots, z_{n_r,r})^\top$, the hyperparameters are chosen by maximizing the log marginal likelihood

$$\mathcal{L}_r(\ell, \sigma_n) := -\frac{1}{2} \mathbf{z}_r^\top \mathbf{A}_{\ell,\sigma_n,r}^{-1} \mathbf{z}_r - \frac{1}{2} \log\det \mathbf{A}_{\ell,\sigma_n,r} - \frac{n_r}{2} \log(2\pi). \tag{64}$$

The implementation searches

$$\begin{aligned} \ell &\in \left\{ 10^{-2+2a/9} : a = 0, \dots, 9 \right\}, \\ \sigma_n &\in \left\{ 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2} \right\}. \end{aligned} \tag{65}$$

These grids are fixed across videos, while the maximizing pair $(\widehat{\ell}_r, \widehat{\sigma}_{n,r})$ is selected separately for each video and refinement round.

For frame f, define

$$\mathbf{k}_{f,r} := \left(k_{\widehat{\ell}_r}(t_f, t_{i_1}), \dots, k_{\widehat{\ell}_r}(t_f, t_{i_{n_r}}) \right)^\top, \qquad \widehat{\mathbf{A}}_r := \mathbf{A}_{\widehat{\ell}_r, \widehat{\sigma}_{n,r}, r}. \tag{66}$$

The estimated score is the posterior mean mapped back to the original score scale:

$$\widehat{y}_{f,r} := \overline{y}_r + s_r \mathbf{k}_{f,r}^\top \widehat{\mathbf{A}}_r^{-1} \mathbf{z}_r. \tag{67}$$

Cached scores are retained at observed frames, so the dense score used by QSTAR is

$$\widetilde{y}_{f,r} := \begin{cases} y_{a,r}, & f = i_a \text{ for some } a \in \{1, \dots, n_r\}, \\ \widehat{y}_{f,r}, & f \notin \mathcal{O}_r. \end{cases} \tag{68}$$

The matrices are factorized by Cholesky decomposition in double precision. If $n_r = 1$ or $s_r < 10^{-6}$, every unobserved frame is assigned the constant score $\overline{y}_r$. The sampler uses the posterior mean only; posterior variance is not used.

C A CLIP-BASED COMPLEMENTARY SAMPLING BRANCH

As one concrete complementary branch, we include a CLIP-based module for cases where the target moment is visually identifiable from the question text.

We first split the video uniformly into $B = 6$ temporal blocks,

$$\mathcal{T}_1, \ldots, \mathcal{T}_B. \tag{69}$$

Within each block, we sample candidate frames at approximately 1 fps. Let the resulting candidate set be

$$\mathcal{I}^{\text{clip}} = \{d_1, \ldots, d_M\}. \tag{70}$$

We encode the question once with the CLIP text encoder and each candidate frame with the CLIP image encoder. The similarity score of candidate d_j is

$$s_j^{\text{clip}} = \left\langle \frac{g_I(f_{d_j})}{\|g_I(f_{d_j})\|_2}, \frac{g_T(X)}{\|g_T(X)\|_2} \right\rangle. \tag{71}$$

To reduce latency, we search block by block rather than score the whole video at once. For block b, let the CLIP scores of its candidates be $\{s_j^{\text{clip}} : d_j \in \mathcal{T}_b\}$. We compute the block mean μ_b, standard deviation σ_b, and maximum score m_b. If

$$m_b > \mu_b + \lambda \sigma_b, \tag{72}$$

we stop exploring later blocks and treat the current block as sufficiently informative. In our implementation, $\lambda = 2.5$. If early stopping is triggered, candidates from unexplored blocks are assigned score zero during final sampling.

The CLIP scores are converted into a sampling distribution

$$p_j^{\text{clip}} = \frac{\tilde{s}_j^{\text{clip}}}{\sum_{\ell=1}^{M} \tilde{s}_\ell^{\text{clip}}}, \tag{73}$$

where $\tilde{s}_j^{\text{clip}}$ is the nonnegative score after deduplication and optional zero-masking of unexplored blocks. We again use ITS without replacement to sample the final K frames:

$$\mathcal{F}^{\text{clip}} = \{f_{d_j} \mid j \in \text{ITS}(p^{\text{clip}}, K)\}, \tag{74}$$

and obtain the CLIP-branch prediction

$$\hat{y}^{\text{clip}} = \text{VLMAnswer}(\mathcal{F}^{\text{clip}}, X). \tag{75}$$

This CLIP branch is then fused with the core saliency-based branch using the output-fusion procedure described in Section 3.4.

D ADDITIONAL ABLATION STUDIES

Ablation across backbone VLMs. To evaluate whether the proposed QSTAR framework generalizes beyond a single backbone, we further apply QSTAR to three additional strong open-source VLMs: InternVL3.5-38B, LLaVA-Video-72B-Qwen2, and QwenVL-2.5-72B. The main results table reports only the strongest configuration, QSTAR with InternVL3-78B, while the remaining backbone variants are reported here as an ablation on the choice of base model.

Table 4 shows that QSTAR consistently provides strong performance across different backbone VLMs, rather than depending on a single model family. In particular, QSTAR with InternVL3.5-38B achieves 81.0 on EgoSchema, 66.7 on LongVideoBench, and 62.9 on MLVU, showing that a smaller 38B backbone can already obtain competitive long-video reasoning performance. Among the 72B-scale backbones, LLaVA-Video-72B-Qwen2 performs strongly on NeXT-QA and Video-MME, while QwenVL-2.5-72B gives the best Video-MME average among the three ablated models. These results suggest that the gains of QSTAR come from the proposed query-adaptive perception and reasoning procedure, and are not tied to one specific backbone architecture.

Table 4: Ablation of QSTAR with different backbone VLMs. Accuracy is reported in percentage. For Video-MME, results are reported in the format w/o subtitle / w. subtitle.

Backbone	NeXT-QA	EgoSchema	LongVideoBench	MLVU	Video-MME			
					Short	Medium	Long	Average
InternVL3.5-38B	84.2	81.0	66.7	62.9	79.4/82.3	66.1/77.1	57.4/72.8	67.7/77.4
LLaVA-Video-72B-Qwen2	84.9	72.4	61.6	56.9	78.6/81.2	66.0/76.7	59.6/73.3	68.0/76.7
QwenVL-2.5-72B	83.9	77.6	64.2	54.9	79.9/82.1	67.7/76.6	60.0/72.8	69.2/77.1

Ablation of Frame Scoring Signals. We further ablate the source of the temporal saliency signal used to guide frame selection. All variants use the same backbone, frame budget, interpolation module, and stopping criterion; only the frame scoring signal is changed. We compare our activation-based score with two alternatives: (i) an attention-based score obtained by averaging the visual-token attention weights within each frame, and (ii) an external CLIP-similarity score computed with EVA-CLIP Radford et al. (2021); Sun et al. (2023). Table 5 shows that the proposed activation signal provides the best overall accuracy–efficiency trade-off.

Table 5: Ablation of different temporal saliency signals. Accuracy is reported in percentage, and inference time is reported as seconds per question. All experiments use InternVL3-78B under the default frame budget.

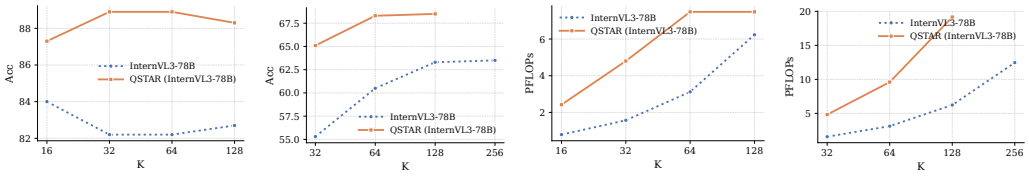
Dataset	Metric	Activation	Attention	CLIP
NeXT-QA	Accuracy ↑	**87.3**	87.2	86.1
NeXT-QA	Time/QA ↓	7s	99s	**3s**
MLVU	Accuracy ↑	**65.1**	64.9	62.9
MLVU	Time/QA ↓	57s	324s	**36s**

On NeXT-QA, attention-based scoring reaches an accuracy very close to activation-based scoring, but it is substantially more expensive, increasing the inference time from 7s/QA to 99s/QA. The CLIP-based scorer is the fastest variant, but its accuracy drops from 87.3% to 86.1%, indicating that shallow image-text similarity is efficient but less aligned with answer-relevant temporal evidence.

The advantage of activation-based scoring becomes more significant on MLVU, where the videos are longer and the answer evidence is more sparse. Activation-based scoring improves over attention-based scoring by 0.2 points and over CLIP-based scoring by 2.2 points. Although CLIP is faster, its external similarity signal cannot reliably capture evidence that only becomes relevant after temporal or compositional reasoning. Attention, in contrast, is both less accurate and much more expensive on MLVU, suggesting that raw attention weights are not a reliable substitute for answer-conditioned activation evidence. Moreover, Attention requires access deeper into the model compared to QSTAR which only needs to read the final-layer hidden states. These results support using activation-based saliency as the default scoring signal in QSTAR. It preserves the semantic grounding of the VLM's own generation process while avoiding the high cost of explicit attention-map aggregation and the weaker reasoning alignment of external CLIP similarity. Recall in Theorem A.4, we analyze the reason why attention-based saliency is much slower. In Theorem A.5, we also show why activation-based saliency is often more accurate.

Effect of frame budget K. Figure 5 shows that simply increasing the number of uniformly sampled frames does not reliably improve the base VLM. On NeXT-QA, the InternVL3-78B baseline drops from 84.0 at $K = 16$ to around 82 at $K = 32$ and $K = 64$, suggesting that more frames can introduce distracting evidence. In contrast, QSTAR improves the same backbone to 87.3 at $K = 16$ and reaches 88.9 at $K = 32$ and $K = 64$. On MLVU, the baseline improves as K increases, but QSTAR is consistently better: with only $K = 32$, QSTAR reaches 65.1, already outperforming the baseline even at $K = 256$. This indicates that the gain comes from selecting more useful frames, not merely from increasing the visual budget.

The cost plots further show that QSTAR has a better accuracy–cost trade-off. Because QSTAR performs iterative perception, its cost at the same K is higher than a single direct VLM call. However, it needs far fewer final frames to reach stronger accuracy. For example, on MLVU, QSTAR with $K = 32$ outperforms direct InternVL3-78B with $K = 256$ while using substantially fewer PFLOPs.



NeXT-QA Acc. MLVU Acc. NeXT-QA Cost MLVU Cost

Figure 5: Ablation on the frame budget K. We compare direct inference with InternVL3-78B and QSTAR using InternVL3-78B.

Similarly, on NeXT-QA, QSTAR with $K = 16$ already exceeds the accuracy of the baseline with much larger frame budgets. These results support the central claim that activation-guided temporal selection is more effective than dense uniform sampling.

Ablation of interpolation methods. Our active sampling procedure requires estimating a dense saliency curve from sparsely observed activation scores. We therefore compare Gaussian process regression (GPR) Williams & Rasmussen (1995) with several commonly used interpolation alternatives: linear interpolation, quadratic interpolation, piecewise cubic Hermite interpolation (PCHIP), and cubic spline interpolation. All variants use the same activation-based frame scores, sampling budget, stopping criterion, and InternVL3-78B backbone; only the temporal interpolation module is changed. As shown in Table 6, the interpolation choice has a larger effect on long-video reasoning than on shorter temporal reasoning. On NeXT-QA, all interpolation methods perform similarly, with GPR and PCHIP both reaching 87.3%. This suggests that when videos are relatively short and the informative evidence is less temporally dispersed, most smooth interpolation schemes can recover a useful sampling distribution. In contrast, the gap becomes clear on MLVU: GPR achieves 65.1%, outperforming linear interpolation by 2.8 points, PCHIP by 2.4 points, spline interpolation by 1.6 points, and quadratic interpolation by 1.4 points. These results indicate that long videos benefit from a more robust global estimator of the saliency curve, since sparse activation scores may be noisy and unevenly distributed across the timeline. We therefore use GPR as the default interpolation module in QSTAR. Its consistently strong performance supports the design choice of using a probabilistic temporal regressor to densify sparse internal saliency observations before updating the sampling distribution.

Table 6: Ablation of interpolation methods for estimating the dense temporal saliency curve. All results use QSTAR with InternVL3-78B under the default frame budget.

Interpolation	GPR	Linear	Quadratic	PCHIP	Spline
NeXT-QA	**87.3**	86.9	86.9	87.3	87.1
MLVU	**65.1**	62.3	63.7	62.7	63.5

Ablation of λ. We further study the stopping threshold λ used in the complementary CLIP-based branch. In this branch, the video is divided into temporal blocks and searched sequentially. For each block, we compute the mean μ_b, standard deviation σ_b, and maximum CLIP score m_b among its candidate frames. The branch stops early when $m_b > \mu_b + \lambda\sigma_b$. Thus, a smaller λ makes early stopping easier and reduces computation, while a larger λ forces the branch to explore more temporal blocks before stopping.

Table 7: Ablation of the CLIP branch stopping threshold λ on LongVideoBench. Accuracy is reported in percentage. #avg explored blocks denotes the average number of temporal blocks evaluated before early stopping.

λ	2.0	2.5	3.0
Accuracy	65.1	66.8	**66.9**
#avg explored blocks	**2.85**	3.96	4.82

Table 7 shows that an overly small threshold hurts accuracy. When $\lambda = 2.0$, the branch explores only 2.85 blocks on average, but the accuracy drops to 65.1%, suggesting that aggressive early stopping may terminate the search before reaching the answer-relevant evidence. Increasing λ to 2.5

substantially improves accuracy to 66.8%, while still requiring fewer than four explored blocks on average.

Using $\lambda = 3.0$ gives the highest accuracy, 66.9%, but the gain over $\lambda = 2.5$ is only 0.1 points while requiring 4.82 explored blocks on average. This indicates a diminishing return from further relaxing the stopping criterion. Therefore, we use $\lambda = 2.5$ as the default setting, since it achieves nearly the same accuracy as $\lambda = 3.0$ with noticeably lower CLIP-search cost.

Ablation of Output Fusion. We further ablate the effect of the output-fusion stage along two dimensions: sampling-method fusion and model fusion. For sampling-method fusion, we evaluate whether adding the complementary CLIP-based branch improves the core saliency-based branch on LongVideoBench. For model fusion, we evaluate whether fusing multiple VLM backbones improves over the best single model on NeXT-QA and MLVU.

Table 8: Ablation of output fusion. Accuracy is reported in percentage. For sampling-method fusion, "w/o Fusion" denotes QSTAR without the CLIP branch, and "w. Fusion" denotes fusion between the core saliency branch and the CLIP branch. For model fusion, "w/o Fusion" denotes the best individual model in the ensemble, and "w. Fusion" denotes the fused result.

Fusion Type	Setting	w/o Fusion	w. Fusion	Gain
	Sampling-method fusion on LongVideoBench			
CLIP branch fusion	InternVL3.5-38B	66.7	**67.6**	+0.9
CLIP branch fusion	InternVL3-78B	65.4	**66.8**	+1.4
CLIP branch fusion	QwenVL-2.5-72B	64.2	**64.8**	+0.6
	Model fusion on NeXT-QA and MLVU			
Model fusion	NeXT-QA: InternVL3-78B + LLaVA-Video-72B-Qwen2 + QwenVL-2.5-72B	87.3	**87.7**	+0.4
Model fusion	MLVU: InternVL3-78B + InternVL3.5-38B + QwenVL-2.5-72B	65.1	**67.9**	+2.8

Table 8 shows that both forms of fusion improve the final accuracy. On LongVideoBench, adding the CLIP-based branch consistently improves all three backbone settings, with gains of 0.6–1.4 points. This supports the motivation for including a complementary CLIP branch: LongVideoBench often contains long videos where the answer depends on a small number of target frames, and the position of these frames is difficult to infer from only a few sampled observations. When the question contains a detailed visual description, CLIP-style image–text matching can directly retrieve relevant frames and provide evidence complementary to the activation-guided saliency branch. We do not include LLaVA-Video-72B-Qwen2 in this setting because it has a shorter effective context length for the fusion discussion prompt.

Model fusion also improves over the best single model. On NeXT-QA, the gain is modest, increasing accuracy from 87.3% to 87.7%, which is expected because the videos are shorter and the best single model is already strong. On MLVU, the gain is larger, improving from 65.1% to 67.9%. This indicates that long-video reasoning benefits more from model diversity, since different VLMs may fail on different temporal, visual, or reasoning cases. The fusion stage can therefore resolve conflicts among model-specific predictions and select a more reliable final answer.

Ablation of CLIP-Style Encoders. We further study the effect of the vision–language encoder used in the complementary CLIP-based branch. The experiments are conducted on LongVideoBench with InternVL3-78B as the backbone VLM. We compare three CLIP-style encoders: the original CLIP Radford et al. (2021), EVA-CLIP Sun et al. (2023), and SigLIP Zhai et al. (2023). All variants use the same temporal block search procedure, stopping threshold, and downstream VLM reasoning module; only the frame–text similarity encoder is changed.

Table 9: Ablation of CLIP-style encoders in the complementary CLIP-based branch on LongVideoBench. Accuracy is reported in percentage, and speed is measured as seconds per frame for similarity scoring. All experiments use InternVL3-78B as the backbone VLM.

Encoder	CLIP	EVA-CLIP	SigLIP
Accuracy ↑	66.5	**66.8**	65.6
s/frame ↓	**0.09**	0.14	0.19

Table 9 shows that the choice of CLIP-style encoder has a moderate but non-negligible impact on the complementary branch. EVA-CLIP achieves the best accuracy, improving over CLIP by 0.3 points and over SigLIP by 1.2 points. This suggests that stronger image–text representations can provide slightly better temporal evidence retrieval for LongVideoBench, where the question often contains explicit textual cues that can be matched to visually grounded events.

However, the accuracy gain of EVA-CLIP comes with additional retrieval cost. Compared with CLIP, EVA-CLIP increases the similarity-scoring time from 0.09s/frame to 0.14s/frame, while SigLIP is both slower and less accurate in this setting. Therefore, EVA-CLIP provides the best accuracy-oriented configuration, whereas CLIP offers the most efficient alternative with only a small accuracy drop. In our default setting, we use EVA-CLIP for the complementary branch because it gives the highest LongVideoBench accuracy while keeping the retrieval overhead small relative to the downstream VLM inference cost.

Ablation of Stopping Condition. We compare QSTAR's KL-divergence-based stopping condition with two alternatives: a fixed number of refinement rounds ($r = 3$ or $r = 5$) and confidence-based stopping with threshold $\theta = 0.8$. Our default criterion terminates refinement when the KL divergence between successive temporal sampling distributions falls below τ. Table 10 reports accuracy and the average number of model calls per question. On NeXT-QA, all stopping rules achieve similar accuracy, with KL-based stopping matching fixed $r = 3$ while requiring 25% fewer calls. The differences are more pronounced on MLVU: KL-based stopping improves accuracy over fixed $r = 3$ by 1.2 percentage points while using fewer calls, and exceeds confidence-based stopping by 2.1 points with only 0.6 additional calls. Increasing the fixed budget to $r = 5$ yields slightly higher accuracy, but KL-based stopping requires 50% fewer calls on NeXT-QA and 38.3% fewer calls on MLVU, with accuracy differences of only 0.1 and 0.4 points, respectively. These results support monitoring sampling-distribution stability as an effective way to balance reasoning accuracy and inference cost.

Table 10: Ablation of stopping conditions on NeXT-QA and MLVU. We report accuracy (%) and average model calls per question.

Stopping condition	NeXT-QA		MLVU	
	Accuracy ↑	Calls ↓	Accuracy ↑	Calls ↓
KL, default τ (ours)	87.3	3.0	65.1	3.7
Fixed $r = 3$	87.3	4.0	63.9	4.0
Fixed $r = 5$	87.4	6.0	65.5	6.0
Confidence $\theta = 0.8$	87.1	2.5	63.0	3.1

Runtime and Refinement Overhead. To further characterize the computational overhead of QSTAR, we report the average frame usage, refinement rounds, VLM calls, and latency breakdown in Table 11. Here, K denotes the *per-iteration* frame budget rather than the total number of frames processed throughout the full procedure. Under the default KL-based stopping criterion, QSTAR performs only 2.00 and 2.73 scoring rounds on average on NeXT-QA and MLVU, respectively, corresponding to 3.00 and 3.73 total VLM calls after including the final answer-generation call. Although the videos contain an average of 1,141 and 26,541 frames, QSTAR scores only 25.7 and 81.2 unique frames during refinement. Moreover, GPR fitting introduces negligible overhead (0.01s and 0.04s), with most runtime spent on VLM scoring and final answer generation. The resulting end-to-end latency is 7s on NeXT-QA and 57s on MLVU. We note that Table 3 compares methods at their default accuracy–efficiency operating points: QSTAR uses $K = 16/32$, whereas the direct VLM uniformly samples at 2 fps with at most 256 frames. Figure 5 separately provides the comparison when the direct VLM uses the same frame budget K as QSTAR.

Comparison with recent structured baselines. We further compare QSTAR with recent methods Liu et al. (2025); Ye et al. (2025); Zhu et al. (2026) that follow a similar select/retrieve-then-reason structure under comparable per-call visual budgets, as summarized in Table 12. We set the number of frames per VLM call for those baseline methods the same as QSTAR and the backbone VLM of compared methods are all set as InternVL3-78B for fair comparison, and for remaining settings we use the default configuration as released in the paper's repository. On NeXT-QA, QSTAR processes 16 frames per VLM call and views only 48 frames in total, compared with 55.1 for BOLT, 56.8 for FOCUS, and 68.2 for T*. QSTAR reaches 87.3% accuracy, exceeding BOLT by 2.1 points and T* by

Table 11: Per-question computational statistics of QSTAR under the default KL-based stopping criterion. K denotes the per-iteration frame budget, and "unique frames scored" counts distinct frames processed during refinement.

Per-question statistic	**NeXT-QA**	**MLVU**
Frames per iteration (K)	16	32
Total frames in video (F, avg.)	1,141	26,541
Scoring rounds (avg.)	2.00	2.73
Total VLM calls (rounds + final QA)	3.00	3.73
Unique frames scored	25.7	81.2
Scoring-pass latency (s)	4.9	41.8
GPR-fitting latency (s)	0.01	0.04
Final-answer latency (s)	2.1	13.9
Total latency (s)	**7**	**57**

1.7 points, while viewing 12.9% and 29.6% fewer total frames, respectively. On the substantially longer MLVU benchmark, QSTAR achieves 65.4% accuracy with only 119.36 total viewed frames, compared with 63.9%/965 frames for BOLT, 60.2%/441 frames for FOCUS, and 57.8%/456 frames for T*. This corresponds to accuracy gains of 1.5, 5.2, and 7.6 points while reducing total viewed frames by 87.6%, 72.9%, and 73.8%, respectively. Due to the usage of strong base VLM itself, QSTAR is able to improve the performance significantly, while bringing a comparable or much lower latency.

Table 12: Comparison with recent methods using a similar select/retrieve-then-reason structure. "VLM frames" counts frames directly processed by the VLM, while "helper frames" counts additional frames processed by auxiliary retrieval or selection modules. "Total viewed" is their sum. All numbers are measured per question.

Method	Accuracy (%)	Time/QA	VLM calls	VLM frames	Helper frames	Total viewed
NeXT-QA						
QSTAR	**87.3**	7.0s	3.00	48	0	**48**
BOLT Liu et al. (2025)	85.2	2.8s	1.00	16	39.1	55.1
FOCUS Zhu et al. (2026)	84.6	4.6s	1.00	16	40.8	56.8
T* Ye et al. (2025)	85.6	6.0s	2.00	24	44.2	68.2
MLVU						
QSTAR	**65.4**	57.0s	3.73	119.36	0	**119.36**
BOLT Liu et al. (2025)	63.9	30.6s	1.00	32	933	965
FOCUS Zhu et al. (2026)	60.2	88.9s	1.00	32	409	441
T* Ye et al. (2025)	57.8	53.9s	2.00	40	416	456

E PSEUDOCODE

Algorithm 1 summarizes the full procedure. The first part is the core QSTAR method. Then in case there are other available sampling approaches, we can do fusion.

F FAILURE CASES

Figure 6 shows a representative failure case on LongVideoBench. The video lasts 8 minutes and 14 seconds, while the target evidence appears for only about 1.5 seconds. This creates an extremely sparse-evidence setting, where the relevant segment occupies only a very small portion of the full video. If the target frame is not sampled, the activation-based perception loop has little direct evidence for identifying where the model should focus.

This failure is also related to the weak temporal continuity of scenes in LongVideoBench. Many videos contain loosely connected scenes, so observing other sampled frames does not reliably indicate where the target event should occur. As a result, when the initial samples miss the target segment, the learned saliency distribution may be guided by visually salient but irrelevant content. In this example, uniformly sampling 32 frames gives only about a 9% chance of covering the target frame. Although the target frame is sampled in the fourth round, the signal is unstable and does not persist in later

Algorithm 1 QSTAR for Long-Video Understanding

Require: video V, question X, frame budget K
Ensure: predicted answer $\hat{y}$

Initialize the sampling distribution p to be uniform distribution over V
for $r = 1, \ldots, R_{\max}$ **do**
  $\mathcal{J} \leftarrow \mathrm{ITS}(p, K)$
  $\mathcal{F}_{\mathrm{tmp}} \leftarrow \{g_j \mid j \in \mathcal{J}\}$
  Run VLM on $\mathcal{F}_{\mathrm{tmp}}$ and the activation-scoring prompt
  Compute sampled frame scores using Section 3.2
  Cache scores on sampled frame indices
  Fit 1D GP on cached scores and predict dense scores $\hat{s}$ over V
  $p_{\mathrm{new}} \leftarrow \mathrm{Normalize}(\max(\hat{s}, \epsilon))$
  if $r > 1$ and $D_{\mathrm{KL}}(p_{\mathrm{new}} \| p) < \tau$ **then**
    $p \leftarrow p_{\mathrm{new}}$
    break
  end if
  $p \leftarrow p_{\mathrm{new}}$
end for
$\mathcal{F}^{(0)} \leftarrow \{g_j \mid j \in \mathrm{ITS}(p, K)\}$
$\hat{y}^{(0)} \leftarrow \mathrm{VLMAnswer}(\mathcal{F}^{(0)}, X)$
$R^{(0)} \leftarrow \mathrm{Reason}(\mathcal{F}^{(0)}, X, \hat{y}^{(0)})$
if no complementary sampling branch is enabled **then**
  return $\hat{y}^{(0)}$
end if
for each enabled complementary branch $m = 1, \ldots, M$ **do**
  Run branch m on (V, X, K)
  Obtain sampled frames $\mathcal{F}^{(m)}$, answer $\hat{y}^{(m)}$, and reason $R^{(m)}$
end for
if all enabled branches predict the same valid answer **then**
  return that answer
end if
if all branches use the same VLM **then**
  $\hat{y} \leftarrow \mathrm{Judge}\Big(X, \big\{(\mathcal{F}^{(m)}, \hat{y}^{(m)}, R^{(m)})\big\}_{m=0}^{M}\Big)$
  return $\hat{y}$
end if
for each branch $i = 0, \ldots, M$ **do**
  for each branch $j = 0, \ldots, M$ **do**
    $s_{i \to j} \leftarrow J_i\Big(X, \big\{(\mathcal{F}^{(m)}, \hat{y}^{(m)}, R^{(m)})\big\}_{m=0}^{M}, j\Big)$
  end for
end for
for each branch $j = 0, \ldots, M$ **do**
  $S_j \leftarrow \sum_{i=0}^{M} s_{i \to j}$
end for
$j^* \leftarrow \arg\max_{j \in \{0, \ldots, M\}} S_j$
return $\hat{y}^{(j^*)}$

sampling rounds. Consequently, the final selected frames miss the target segment again, leading the model to answer based on non-target evidence. This case highlights a limitation of activation-guided temporal refinement when the decisive evidence is extremely short and cannot be inferred from the surrounding sampled scenes.

Figure 7 shows how the CLIP branch fixes the failure in Figure 6. Because the target event is very short, the activation-guided branch may miss the target frame and answer based on irrelevant evidence. In contrast, the CLIP branch performs direct query-frame matching and identifies the region most relevant to the question. In this example, it successfully retrieves frames containing the target dog and produces the correct answer "a dog," while the activation-guided branch predicts "a man." The final

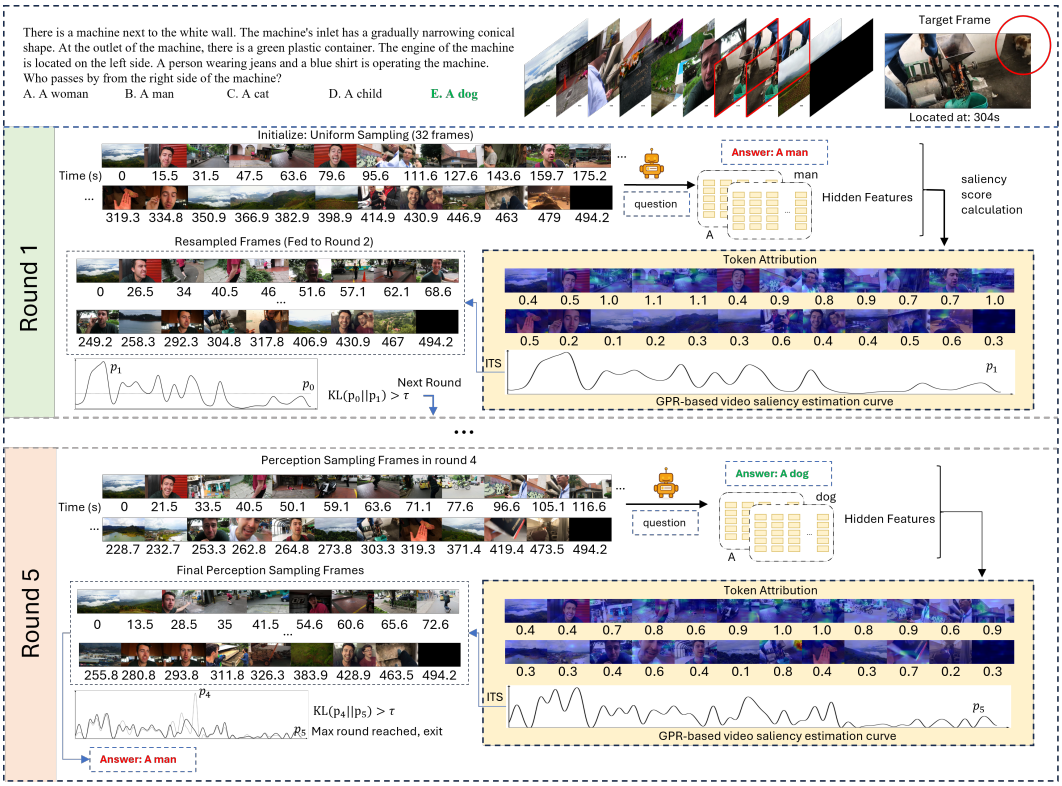


Figure 6: Failure case under extremely sparse temporal evidence on LongVideoBench.

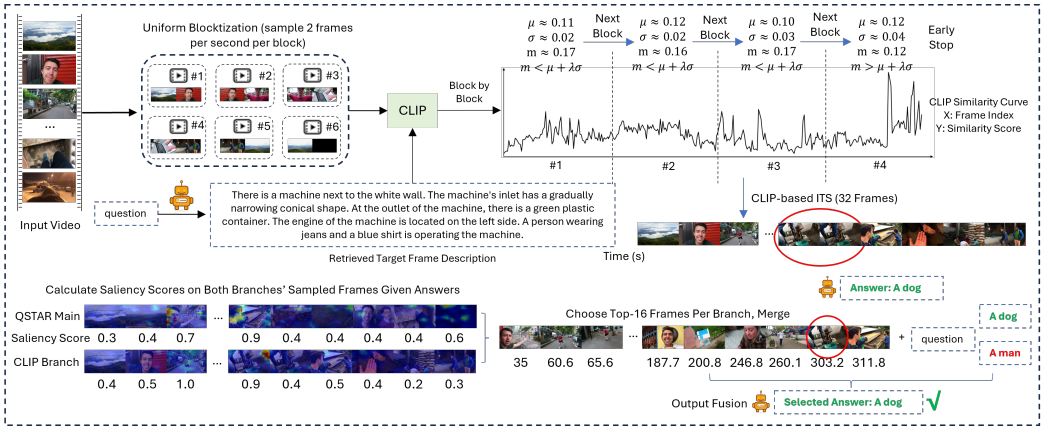


Figure 7: CLIP branch recovery for the sparse-evidence failure case.

fusion step then selects the CLIP-supported answer. This case shows that CLIP is especially helpful when the target evidence is brief but visually specific. However, we should also note that CLIP may not succeed on each such case. For example, if the question contains temporal relationships like "before" or "after" a described scene, then CLIP is unable to process this kind of relationship. In such cases, we may need to combine with neural-symbolic methods, and we consider it as our future work.

Figure 8 shows another failure case from MLVU with a different cause. Here, the relevant frames are successfully sampled, so the error is not caused by temporal perception. Instead, the backbone model

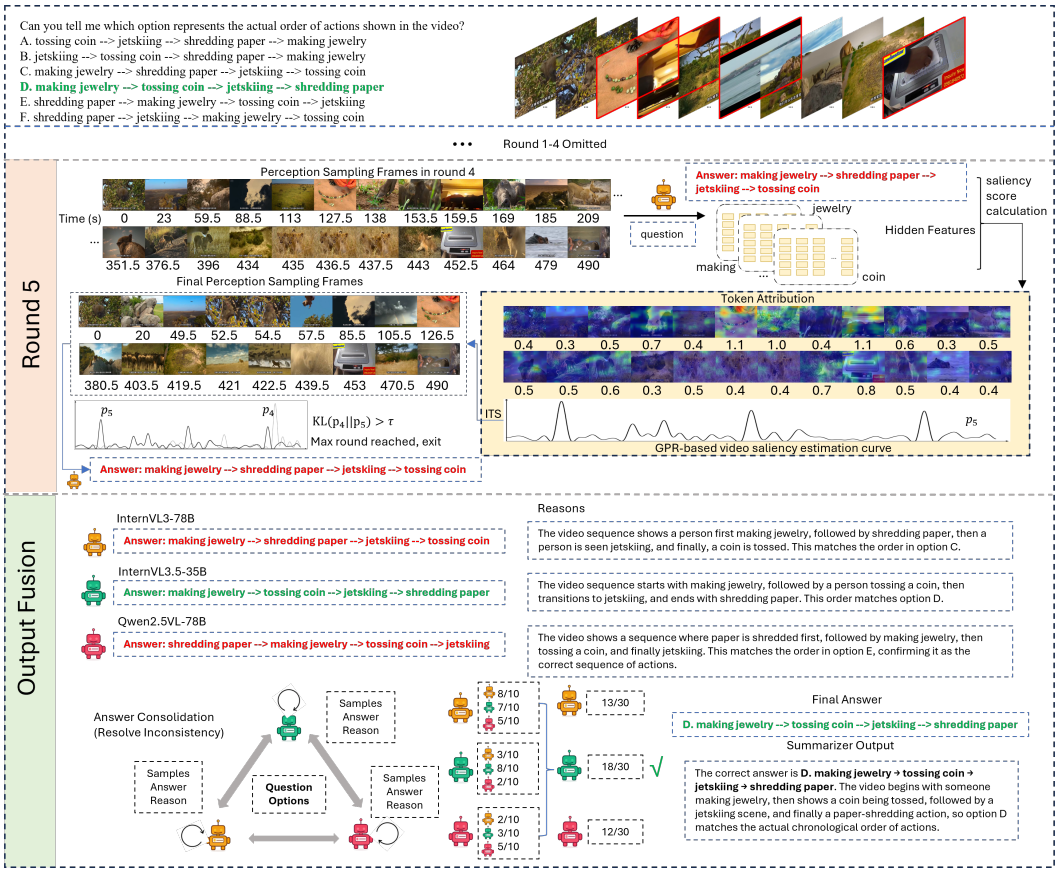


Figure 8: Recovery from backbone answer failure by model fusion.

InternVL3-78B fails to correctly recognize the "shredding paper" event, which leads to an incorrect prediction even though the key evidence is already available.

The lower part of Figure 8 shows how output fusion resolves this issue. Although the three agents produce different answers, the agent that predicts the correct choice receives an overwhelmingly higher score after cross-model discussion and evaluation. As a result, the final fused prediction is correct.

Figure 9 shows a failure case on NeXT-QA. In this video, many sampled frames are near-duplicates, and the correct answer is "lay still." However, the backbone model predicts "move his head around." Since QSTAR's activation score estimates how much each frame explains the generated answer, an incorrect answer can produce misleading frame scores when the visual content is highly repetitive. As a result, the sampling distribution may be pushed toward irrelevant duplicated frames instead of correcting the initial mistake. Even if we use model fusion, the mistake is not resolved.

This example clarifies the role of the activation-based score. The score measures answer-conditioned visual support: it identifies frames whose visual representations support the provisional answer tokens produced by the VLM. This conditioning makes the score specific to the model's current reasoning state. When the provisional answer is incorrect and the video contains many near-duplicate frames, visually similar frames may receive comparable support and remain prominent across refinement rounds. One extension is to compute contrastive scores over multiple answer hypotheses, so that frame selection reflects evidence distinguishing the alternatives rather than support for a single provisional answer. The optional complementary branch provides an additional source of evidence in such cases.

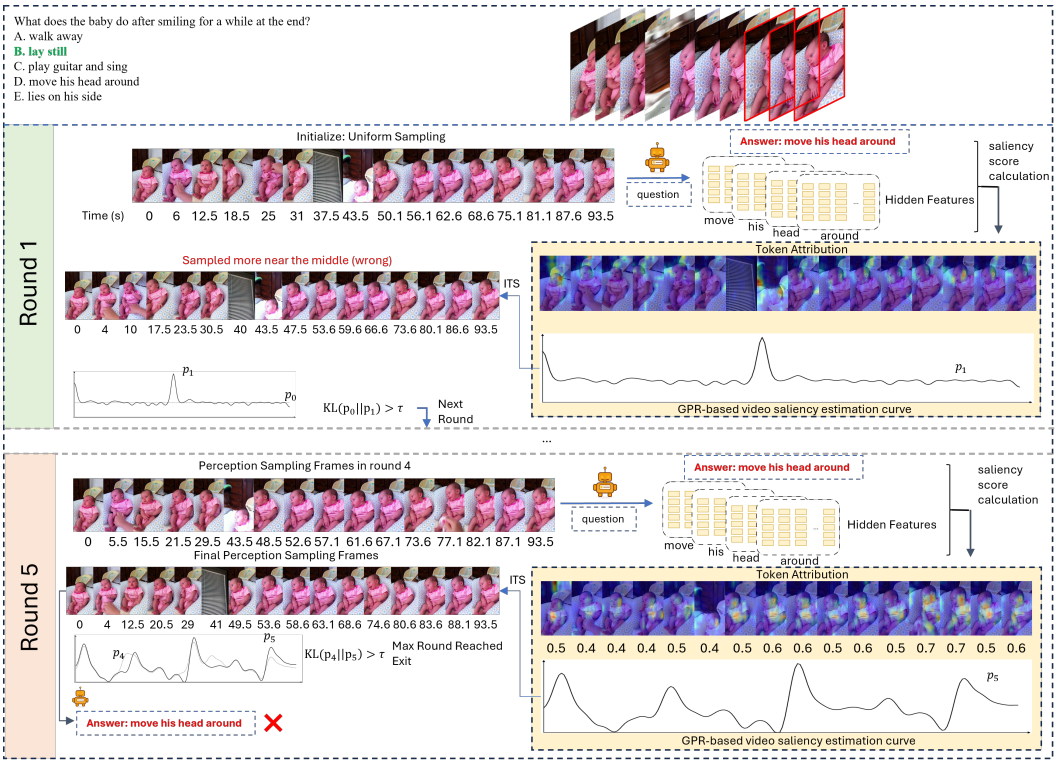


Figure 9: Failure case caused by duplicated frames on NeXT-QA.

Table 13: Answer transitions from the first scoring round to the final prediction of the core QSTAR saliency branch. Percentages for the first two rows are computed over initially incorrect predictions, while percentages for the next two rows are computed over initially correct predictions.

Answer transition	NeXT-QA	MLVU
Wrong → Correct	53 / 205 (25.9%)	64 / 197 (32.5%)
Wrong → Wrong	152 / 205 (74.1%)	133 / 197 (67.5%)
Correct → Wrong	39 / 1,288 (3.0%)	34 / 292 (11.6%)
Correct → Correct	1,249 / 1,288 (97.0%)	258 / 292 (88.4%)
Net corrections	+14	+30
Net accuracy change	+0.9 pp	+6.1 pp

Quantifying answer correction and error reinforcement. To quantify how often iterative refinement corrects or reinforces an initially incorrect prediction, we compare the answer generated during the first scoring round with the final prediction of the core saliency branch, before applying CLIP-based or model fusion. Both predictions use the same backbone, answer prompt, decoding procedure, and per-call frame budget; only the selected frames change. Table 13 reports the resulting answer transitions.

On NeXT-QA, refinement corrects 53 of the 205 initially incorrect predictions (25.9%), while changing 39 of the 1,288 initially correct predictions to incorrect (3.0%), resulting in 14 net corrections. On MLVU, it corrects 64 of 197 initially incorrect predictions (32.5%) and changes 34 of 292 initially correct predictions to incorrect (11.6%), resulting in 30 net corrections. Thus, the refinement process produces more corrections than regressions on both benchmarks, although answer-conditioned error reinforcement remains non-negligible, particularly on MLVU. This limitation motivates future extensions based on contrastive evidence scoring over multiple answer hypotheses.

G LIMITATIONS

Although QSTAR achieves strong results, it still has several limitations. First, the active perception process may fail when the decisive evidence is very short and is missed in early sampling rounds. This is especially difficult for long videos with weakly connected scenes, where the target location cannot be reliably inferred from other sampled frames.

Second, although our method doesn't need a fully white-box access to the VLM, it still requires the final hidden state to compute the saliency score. Therefore, for some close-sourced VLMs whose final hidden state is not accessible via APIs, other techniques are needed to dig out the frame importance.

Third, QSTAR depends on the recognition ability of the backbone VLM. Even if the relevant frames are sampled, the model may still fail to recognize fine-grained actions or objects. Model fusion can alleviate this issue, but it cannot fully remove errors shared across models.

H AI USE STATEMENT

This work uses large vision-language models (VLMs) as core inference components of QSTAR. Given sampled video frames and a question prompt, the VLM generates intermediate or final answers and provides internal hidden states used for activation-based frame-saliency estimation. The proposed method does not train, fine-tune, or modify the parameters of these models; instead, it treats them as fixed inference backbones and uses their generated outputs and internal activation signals to guide query-adaptive frame selection. In the action fusion stage, VLMs may also be used as branch evaluators to compare candidate answers produced by different sampling branches and select the final prediction.

I ETHICS STATEMENT

This work studies efficient long-video understanding using existing video question-answering benchmarks and does not involve new human-subject data collection. QSTAR relies on pretrained vision–language models and may therefore inherit biases, recognition errors, and other limitations of the underlying models and evaluation datasets. As discussed in the limitations, the method may also miss sparse evidence or produce incorrect predictions when the backbone model fails to recognize relevant visual content. These limitations should be considered when applying the method in real-world or high-stakes settings.

J REPRODUCIBILITY STATEMENT

We provide detailed descriptions of QSTAR's activation-based frame scoring, adaptive sampling, stopping criterion, and action fusion in the main paper. Experimental settings, frame budgets, benchmark protocols, and hardware configurations are reported in Section 4. The appendix provides the theoretical assumptions and proofs, implementation details, pseudocode, hyperparameter settings, extensive ablations, runtime analysis, and failure cases. We also release anonymized source code at `https://anonymous.4open.science/r/QSTAR` to facilitate reproduction of the reported results.