NAVIAGENT: COARSE-TO-FINE ACTIVE FRAME SELECTION FOR LONG-FORM VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long videos contain vast temporal search spaces, yet answer-critical visual evidence may be confined to a few brief moments. Under limited frame budgets, effectively locating such sparse evidence requires both broad temporal coverage and fine-grained temporal precision. This makes it challenging to search the long temporal horizon efficiently while preserving the resolution needed to capture decisive evidence. To address this challenge, we propose **NavIAgent**, a coarse-to-fine active frame selection framework that combines global temporal localization with fine-grained evidence acquisition. NavIAgent operates through two complementary stages. (1) At the coarse stage, a lightweight multimodal event graph indexes structured segment memories and rapidly localizes query-relevant candidate intervals. (2) At the fine stage, we deploy **S**trategic **M**emory **I**nformed **L**ens for **E**vidence (Smile), an active agent optimized via reinforcement learning with dual-scale advantage and Free Energy-inspired intrinsic rewards. By formulating keyframe selection as a Partially Observable Markov Decision Process (POMDP), Smile performs discrete temporal navigation within candidate intervals and progressively acquires answer-relevant evidence from intermediate observations. Theoretical analysis shows that active policies have a more expressive hypothesis space than passive sampling, while unconstrained discrete navigation suffers from asymptotic temporal resolution dilution as video length increases, providing a formal motivation for our coarse-to-fine design. Extensive experiments on Video-MME, LongVideoBench, and LVBench across multiple vision-language backbones demonstrate consistent accuracy improvements of up to **11.6** percentage points, while substantially reducing visual frame consumption and inference latency. Our code is in the supplementary materials.

1 INTRODUCTION

Long-form video comprehension poses a significant computational challenge. Processing videos containing tens of thousands of frames can easily exceed the context-window and memory capacities of modern vision-language models (VLMs) (He et al., 2026a; Geng et al., 2025; Tan et al., 2025; Ma et al., 2025; Zhu et al., 2026; Jiang et al., 2025). To make such videos tractable, a common strategy is to select a compact subset of informative frames as visual evidence for answering video-related queries. Previous methods typically rely on heuristic selection strategies, such as query-aware selection (Liu et al., 2025; Peng et al., 2026), adaptive and coverage-aware frame sampling (Tang et al., 2025; Zou et al., 2026b; Zhang et al., 2026c), and grounding-based frame picking (Yu et al., 2025; Tao et al., 2026; Wei et al., 2026; Chen et al., 2026a), focusing on the semantic similarity between the query and frames but ignoring how the selected frames improve the final decision and containing lots of redundant similar frames.

This limitation has driven a shift toward active video perception, which recasts keyframe selection as an iterative, agentic decision-making process. Rather than relying on one-shot heuristic filtering, an active perception agent progressively interacts with the video environment by evaluating current reasoning states and adaptively acquiring new frames from informative temporal regions. For example, recent works learn adaptive acquisition policies (Zou et al., 2026a), enable multi-turn interrogation to query visual evidence on demand (He et al., 2026b), or leverage agentic loops to iteratively locate and verify critical moments (Li et al., 2025; Liu et al., 2026b). By allowing intermediate observations to guide subsequent exploration, these approaches move beyond static selection.

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.

However, the existing methods of active perception still face challenges. The agent must search a large temporal space while precisely capturing short answer-critical events, creating a tradeoff between global coverage and local temporal precision[1], as illustrated in Figure 1(a). Meanwhile, active navigation progressively accumulates answer-relevant evidence, yet existing approaches do not explicitly model how intermediate observations improve evidence sufficiency during navigation. As a result, the agent may acquire informative observations without knowing whether they meaningfully advance toward sufficient evidence, as illustrated in Figure 1(b). In short, current active agents face an unresolved tradeoff between long-range search and fine-grained precision, coupled with an unguided evidence accumulation process.

To address these challenges, we propose NavIAgent, a coarse-to-fine active frame selection framework with two complementary components, as illustrated in Figure 1(c). (*i*) Our Smile addresses evidence-seeking by formulating keyframe selection as a Partially Observable Markov Decision Process (POMDP) and learning temporal navigation with a dual-scale advantage, enabling intermediate observations to guide subsequent actions toward sufficient evidence. (*ii*) The multimodal event graph addresses the large temporal search space by organizing the video into structured segment memories and localizing relevant candidate intervals before active navigation. Together, these components separate global evidence localization from local evidence acquisition, enabling active perception to maintain both efficient long-range search and fine-grained temporal precision.

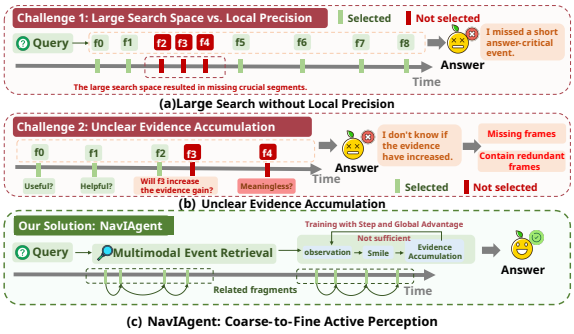

Figure 1: **Overview of long-video perception paradigms. (a) Large search space** makes it difficult to capture short answer-critical events. **(b) Unclear evidence accumulation** makes it difficult to determine whether intermediate observations advance toward sufficient evidence. **(c) Coarse-to-fine active perception** first localizes relevant temporal regions and then performs closed-loop evidence seeking within them.

Our main contributions are summarized as follows:

- **Active Perception Agent (Smile):** We formulate keyframe selection as a partially observable Markov decision process (POMDP) and develop Smile, a closed-loop temporal navigation agent. Trained via reinforcement learning with dual-scale advantage estimation and Free-Energy-inspired intrinsic rewards, Smile adaptively executes percentage-based jumps within candidate segments to achieve precise, frame-level visual evidence localization.

- **Coarse-to-Fine Architecture (NavIAgent):** We introduce NavIAgent, a synergistic framework that unifies multimodal event graph indexing with localized active sensing. Operating under an index-once and query-repeatedly paradigm, the event graph captures cross-segment causal progressions and entity coreferences to prune the search space without digesting raw frame tokens, thereby addressing both the observation-selection divorce of passive sampling and the resolution dilution of unconstrained temporal navigation.

- **Theoretical Analysis & Empirical Evaluation:** We theoretically prove that active perception subsumes passive sampling in expressive power, while formally establishing the asymptotic resolution dilution in discrete temporal navigation. Empirically, extensive evaluations on Video-MME, LongVideoBench, and LVBench demonstrate that NavIAgent enhances base vision-language models by up to 11.6% in accuracy.

[1] *Global coverage* refers to exploring the entire video timeline to locate relevant intervals, whereas *local temporal precision* denotes pinpointing frame-level critical moments within those intervals.

2 Related Work

2.1 Active Perception and Sequential Reinforcement Learning

Recent work applies Reinforcement Learning (RL) and agentic reasoning to temporal video understanding. Video-R1 (Feng et al., 2026), VideoRFT (Wang et al., 2026a), and DeepVideo-R1 (Park et al., 2026) use RL to improve temporal reasoning and temporal localization, while Video-Thinker (Wang et al., 2025a), ReWatch-R1 (Zhang et al., 2026b), and LongVILA-R1 (Chen et al., 2026b) explore interleaved reasoning, visual re-examination, and long-context scaling. Beyond reasoning over pre-selected inputs, active perception methods directly control the acquisition of visual observations. Active Video Perception (Wang et al., 2026c), OmniAgent (Xing et al., 2026), and SAGE (Jain et al., 2026) enable sequential perception through iterative exploration and decision making. However, these active approaches typically navigate the global timeline with relative strides, which can dilute temporal resolution and miss transient evidence as video duration grows.

2.2 Query-Aware Keyframe Selection

Query-aware selection methods extract compact, question-relevant visual subsets to reduce redundant computation and focus perception on relevant content. T* (Ye et al., 2025) and LENS (Zhang et al., 2026a) adaptively narrow temporal or spatial regions, while Flexible Frame Selector (Buch et al., 2025) and K-frames (Yao et al., 2025) dynamically adjust frame budgets according to video content. FocusGraph (Zemskova et al., 2026) further exploits graph structures to reduce visual redundancy, while QGate (Wang et al., 2026b) and TCS (Tan et al., 2026) investigate dynamic modality routing and hierarchical sampling. These methods, however, generally select visual subsets in an open-loop manner, without iterative verification or evidence-driven temporal adaptation after new observations are acquired.

3 Method

In this section, we present NavIAgent, a coarse-to-fine active frame selection framework for long-form video question answering. Formulated as a Partially Observable Markov Decision Process (POMDP) and grounded in theoretical analysis (Sec.3.1), NavIAgent decouples evidence discovery into two complementary stages to balance global temporal coverage with local resolution: (*i*) a coarse stage (Figure 2(A)) that constructs a multimodal event graph to retrieve candidate intervals without processing dense visual tokens (Sec.3.2); and (*ii*) a fine stage (Figure 2(B)) where Smile, an active navigation agent, executes closed-loop temporal jumps strictly within the retrieved intervals to isolate frame-level evidence (Sec.3.3). Finally, the multimodal evidence (Figure 2(C)) is aggregated for downstream question answering (Sec.3.4).

3.1 Problem Formulation and Theoretical Rationale

Designing an effective frame selection framework for long-form videos hinges on resolving two theoretical questions: (*i*) **The necessity of active perception:** Does an active, iterative perception strategy genuinely outperform conventional passive sampling in capturing sparse, answer-critical evidence? (*ii*) **The inadequacy of unconstrained navigation:** Even with active perception, why does unconstrained global navigation suffer from severe efficiency bottlenecks, and how does this necessitate a coarse-to-fine paradigm?

In this section, we formally address these two questions. By establishing the representational limits of passive sampling and the asymptotic resolution dilution of global navigation, we provide the mathematical rationale that directly motivates the synergistic design of NavIAgent.

(1) Overcoming Representational Limits via Active Perception. To answer the first question, we cast active keyframe selection as a Partially Observable Markov Decision Process with tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O}, \mathcal{Z}, \gamma \rangle$ over a video V comprising N frames. The latent state $s_t \in \mathcal{S}$ represents the underlying multimodal content of frame f_t at interaction step t. The agent does not observe s_t directly, but receives a distilled observation $o_t = \mathcal{Z}(f_t) \in \mathcal{O}$ instantiated as a descriptive textual summary d_t. The navigation operator $\mathcal{T}$ maps $\mathcal{S} \times \mathcal{A} \to \mathcal{S}$, transitioning the current frame index to the next frame via percentage-based relative jumps $a_t \in \mathcal{A}$ that mimic dragging a progress bar. A

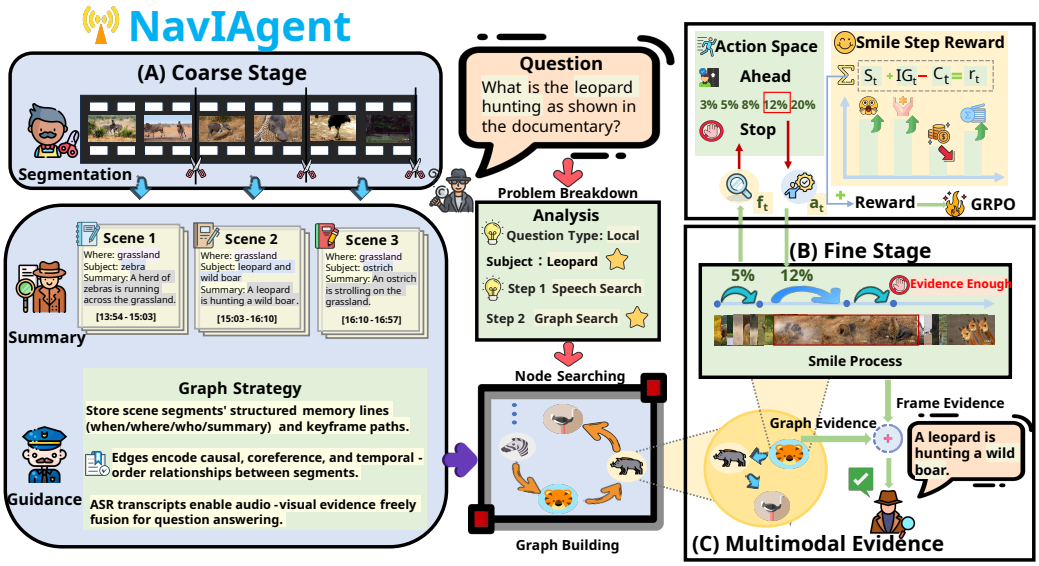

Figure 2: **Overview of the NavIAgent framework. (A) Coarse Stage:** Constructs a multimodal event graph storing structured scene memories to rapidly filter candidate temporal intervals based on problem breakdown. **(B) Fine Stage:** Smile navigates candidate segments via discrete relative jumps (a_t), guided by a step reward and optimized with GRPO. **(C) Multimodal Evidence:** Downstream QA combines graph context and frame-level visual evidence for final prediction.

persistent textual memory $\mathcal{M}_t$ accumulates descriptions of visited frames alongside query context to represent internal state. Given question Q with ground-truth answer $y^* \in \mathcal{Y}$ and exploration budget T, the agent iteratively samples action $a_t \sim \pi(\cdot \mid \mathcal{M}_{t-1}, Q)$, transitions to frame f_{t+1}, and updates memory $\mathcal{M}_t$. This sequential formulation allows subsequent temporal decisions to condition on intermediate evidence. We formally show that this expands the policy hypothesis space in Proposition 1.

Proposition 1 (Representational separation and disambiguation capacity)**.** *For an observation budget $T \geq 2$, the hypothesis class of passive selection policies $\Pi^{\text{pass}}(T)$ is a subset of the active policy class $\Pi^{\text{act}}(T)$*

$$\Pi^{\text{pass}}(T) \subsetneq \Pi^{\text{act}}(T). \tag{1}$$

Proof is in Appendix B.1.

Proposition 1 addresses the first question by establishing the superiority of active perception over passive sampling. Active sensing is mathematically indispensable for resolving observation-dependent ambiguity, providing theoretical justification for designing our active navigation agent.

(2) Resolving the Tradeoff between Coverage and Precision. As highlighted in Sec. 1, the agent must search a large temporal space while precisely capturing short answer-critical events, creating an inherent tradeoff between global coverage and local temporal precision. Navigating an unconstrained global timeline with a fixed budget of point inspections fails to resolve this tradeoff due to asymptotic temporal resolution dilution. Specifically, when navigation actions rely on discrete relative temporal strides to maintain broad coverage, the physical displacement of each action grows linearly with video duration, meaning a fixed budget covers a diminishing fraction of the timeline. We formalize this problem in Proposition 2.

Proposition 2 (Resolution dilution under fixed-budget navigation)**.** *Let L denote the video duration and let an answer-critical event have duration $w_{\text{event}} \ll L$. Suppose its start time τ^* is uniformly distributed over $[0, L - w_{\text{event}}]$, and observations outside the target event provide zero mutual information about τ^* (serving as an idealized, worst-case abstraction). For any active policy with an exploration budget of T point inspections, the probability of missing the target event satisfies*

$$P_{\text{miss}}(L, w_{\text{event}}, T) \geq 1 - \frac{T w_{\text{event}}}{L - w_{\text{event}}}. \tag{2}$$

Consequently, for any fixed T and w_{event},

$$\lim_{L\to\infty} P_{\text{miss}}(L, w_{\text{event}}, T) = 1. \tag{3}$$

Proof is in Appendix B.2.

This mathematical inevitability necessitates restricting the active navigation space to a compact local duration $L_{local} \ll L$, ensuring that the minimum physical displacement $\delta_{min} L_{local}$ remains smaller than the target event duration w_{event}. This is exactly the role of our Coarse Stage.

3.2 Coarse Stage: Multimodal Event-Graph Indexing and Retrieval

Grounded in the above insights, NavIAgent first builds a structured temporal memory to retrieve a compact set of query-relevant candidate segments before activating fine-grained sensing. The detail of segment building is shown in Appendix D.3.

(1) Structured Temporal Memory. We segment the video into M temporally ordered chunks $\mathcal{C} = \{c_1, \ldots, c_M\}$. Each chunk c_i is distilled into a structured memory record:

$$m_i = \mathcal{F}_{\text{mem}}(c_i) = \langle \sigma_i, [t_i^{\text{start}}, t_i^{\text{end}}], \ell_i, \xi_i \rangle, \tag{4}$$

where σ_i is a concise semantic summary, $[t_i^{\text{start}}, t_i^{\text{end}}]$ denotes the temporal boundary, ℓ_i is the spatial environment or scene setting, and ξ_i is the primary subject or entity. These records form the vertices of a multimodal temporal memory graph:

$$\mathcal{G} = (\mathcal{C}, E), \quad E = E_{\text{causal}} \cup E_{\text{coref}} \cup E_{\text{temporal}}, \tag{5}$$

where directed edges link chunks exhibiting causal progressions, shared entity coreferences, or immediate temporal contiguity, facilitating multi-hop relational context expansion. Detail of edge building is shown in Appendix D.3.

(2) Query-Guided Temporal Retrieval. Given Q, a query planner decomposes the user query into retrieval keys $\mathcal{Q}$, which are matched against the structured graph records to identify relevant candidate segments:

$$\mathcal{C}_Q = \text{Expand}_{\mathcal{G}} \left(\bigcup_{q \in \mathcal{Q}} \mathcal{R}(q, \mathcal{G}) \right), \tag{6}$$

where $\mathcal{R}$ retrieves top matching segments and $\text{Expand}_{\mathcal{G}}$ traverses adjacent incident edges in $\mathcal{G}$ to incorporate necessary causal and temporal context. The subset $\mathcal{C}_Q \subseteq \mathcal{C}$ defines a compact search space for subsequent frame-level active exploration.

3.3 Fine Stage: Smile Active Exploration Agent

Within each bounded candidate interval $c_i \in \mathcal{C}_Q$, where the effective temporal duration is bounded to reduce the resolution limitation described in Proposition 2, we deploy Smile to execute closed-loop frame selection.

3.3.1 Overall Execution Flow and Context Consolidation

As illustrated in Figure 3, the execution of Smile follows an iterative *Perceive-Decide-Transition* cycle across steps $t \in \{1, \ldots, T\}$ within the temporal span $[t_i^{\text{start}}, t_i^{\text{end}}]$ of candidate chunk c_i. At step t, the input space is factorized into two decoupled representations: the transient visual percept f_t and the persistent textual memory $\mathcal{M}_{t-1} = \{Q, d_1, \ldots, d_{t-1}\}$, which consolidates the question Q and historical summaries $d_{<t}$. The policy network π_θ first consumes the current raw frame and persistent memory to sample a temporal navigation action, and subsequently distills the frame into a summary description via a secondary generation pass:

$$a_t \sim \pi_\theta^{\text{action}}(\cdot \mid \mathcal{M}_{t-1}, f_t), \quad d_t \sim \pi_\theta^{\text{desc}}(\cdot \mid \mathcal{M}_{t-1}, f_t). \tag{7}$$

This operational cycle decouples into two complementary components:

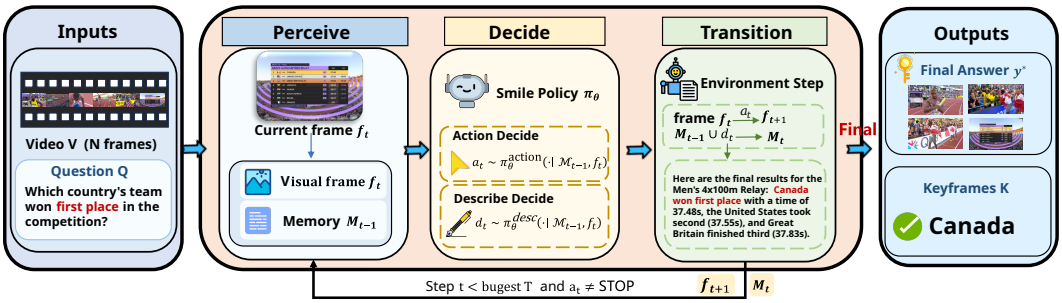

Figure 3: **Execution workflow of Smile.** The agent operates in an iterative loop: perceiving the current visual frame and accumulated textual memory, deciding temporal jumps or visual summaries via π_θ, and updating memory until termination.

- **Temporal Navigation (a_t):** A symbolic operator sampled from the action space $\mathcal{A}$, which is calibrated against the temporal span of the local candidate chunk $[t_i^{\text{start}}, t_i^{\text{end}}]$ rather than the global timeline. (The action taxonomy and discretization details are given in Appendix D.1).
- **Visual Distillation (d_t):** A concise textual description encapsulating task-relevant visual evidence observed in frame f_t.

Upon executing action a_t, the environment advances to frame f_{t+1}. The memory footprint is constrained by discarding the visual tokens of f_t and retaining solely the distilled textual evidence d_t:

$$\mathcal{M}_t = \mathcal{M}_{t-1} \cup \{d_t\}. \tag{8}$$

3.3.2 RL Training via Step-Level Distributed Rewards

Optimizing an active navigation policy over continuous video spaces is hindered by delayed and sparse feedback, as terminal QA accuracy provides insufficient credit assignment for intermediate navigation choices. Inspired by the Free Energy Principle (FEP) Friston (2010), which posits that an adaptive agent must minimize uncertainty while avoiding excessive energy expenditure, we formulate a decomposed step-level distributed reward to supervise intermediate jumps alongside terminal verification.

***(i)* Distributed Per-Step Reward.** At transition step t, the policy receives an immediate composite reward balancing visual exploration, evidence acquisition, and navigation efficiency:

$$R_t = \beta \cdot \|\text{CLIP}_{\text{img}}(f_t) - \text{CLIP}_{\text{img}}(f_{t-1})\|_2 \; + \; \alpha \cdot \max(b_t - b_{t-1},\, 0) \; - \; \lambda \cdot |a_t|, \tag{9}$$

where $\text{CLIP}_{\text{img}}(\cdot)$ denotes the L_2-normalized embedding from a frozen CLIP image encoder. The first term encourages escaping visually redundant frames; the third term penalizes unnecessarily erratic jump magnitudes $|a_t|$ with penalty factor λ; and the second term rewards incremental improvements in the internal evidence belief state b_t. The belief state is updated via exponential smoothing:

$$b_t = \rho \cdot b_{t-1} + (1-\rho) \cdot \min\left(\frac{p_2}{\max(p_1, \varepsilon_{\text{eps}})},\, C\right), \tag{10}$$

where $\rho \in (0,1)$ is a momentum coefficient, $\varepsilon_{\text{eps}} > 0$ prevents division by zero, and $C > 0$ clips belief spikes. Here, p_1 and p_2 quantify the perceived answer sufficiency of the current memory $\mathcal{M}_t$ without and with access to the ground-truth answer y^*, respectively. The exact estimation protocol for p_1, p_2 and hyperparameter configurations are detailed in Appendix D.2.

***(ii)* Trajectory Return and Policy Optimization.** When the agent invokes action `STOP` or exhausts the horizon budget T, the accumulated memory $\mathcal{M}_T$ is utilized to infer the final answer. A terminal reward $R_{\text{term}} \in [0,1]$ is assigned based on semantic equivalence with the ground-truth answer y^*. The total trajectory return G_i for trajectory i of length T_i integrates discounted per-step rewards with terminal verification:

$$G_i = \sum_{t=1}^{T_i} \gamma^{t-1} R_{i,t} + \lambda_{\text{term}} \cdot R_{\text{term},i}, \tag{11}$$

where $\gamma \in [0, 1]$ is the discount factor and λ_{term} balances terminal correctness against intermediate step efficiency.

Dual-scale Advantage. To jointly align global task success with local navigation precision, we construct a dual-scale advantage for trajectory i within a rollout group of size K:

$$A_{i,t} = \underbrace{\frac{G_i - \mu_G}{\sigma_G}}_{\text{Macro}} + \omega \cdot \underbrace{\frac{R_{i,t} - \mu_{\text{step}}}{\sigma_{\text{step}}}}_{\text{Micro}}, \tag{12}$$

where μ_G and σ_G are the empirical mean and standard deviation of total returns across the K parallel rollouts, while μ_{step} and σ_{step} normalize step-level rewards exclusively across active transitions, balanced by ω. The macro component promotes trajectories with high overall returns, while the micro term reinforces fine-grained transitions. The policy is optimized via the clipped GRPO objective:

$$\begin{aligned}\mathcal{J}(\theta) = \mathbb{E}_{\{\tau_i\}_{i=1}^{K} \sim \pi_{\theta_{\text{old}}}} \Bigg[& \frac{1}{T_{\text{total}}} \sum_{i=1}^{K} \sum_{t=1}^{T_i} \min(r_{i,t}(\theta) A_{i,t},\ \text{clip}(r_{i,t}(\theta), 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}) A_{i,t}) \\ & - \beta_{\text{KL}} \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) + \beta_{\text{ent}} \mathcal{H}(\pi_\theta) \Bigg],\end{aligned} \tag{13}$$

where $T_{\text{total}} = \sum_{i=1}^{K} T_i$ denotes the total number of interaction steps across the group, $r_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t}, d_{i,t} | \mathcal{M}_{i,t-1}, f_{i,t})}{\pi_{\theta_{\text{old}}}(a_{i,t}, d_{i,t} | \mathcal{M}_{i,t-1}, f_{i,t})}$ is the importance sampling ratio, and ϵ_{clip} denotes the clipping threshold. Training dynamics and efficiency metrics for Smile are reported in Appendix E.1 and E.2.

3.4 Evidence Aggregation and Answering

The navigation processes across candidate segments run in parallel, and the set of discovered keyframes is aggregated as:

$$\mathcal{K} = \bigcup_{c_i \in \mathcal{C}_Q} \mathcal{N}(c_i, Q), \tag{14}$$

where $\mathcal{N}(c_i, Q)$ denotes the sequence of keyframes selected by Smile within chunk c_i. And the frame descriptions generated during active navigation are collated alongside their parent segment memories and time aligned speech transcripts to construct the final evidence context:

$$\mathcal{W} = \text{Concat}_{c_i \in \mathcal{C}_Q} \Big(m_i \ \circ\ [d_k]_{k \in \mathcal{K}_i} \ \circ\ [u]_{u \in \mathcal{U}_i} \Big), \tag{15}$$

where $\mathcal{K}_i \subseteq \mathcal{K}$ denotes the keyframes visited within chunk c_i, d_k is the textual description of keyframe k, $\mathcal{U}_i$ is the set of aligned speech utterances spanning chunk c_i, and $\circ$ denotes string concatenation. The final answer is generated by prompting the frozen downstream VLM conditioned jointly on the original question Q, the selected visual keyframes $\mathcal{K}$, and aggregated textual context $\mathcal{W}$, formulated as $\text{VLM}(Q, \mathcal{K}, \mathcal{W})$. The training algorithm for Smile is shown in the Appendix C.2.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on three representative long video multiple choice benchmarks: Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and LVBench (Wang et al., 2025b). These benchmarks contain videos of up to two hours and require locating sparse, task-relevant visual evidence over long temporal horizons. The navigation policy Smile is trained exclusively on NExT-QA (Xiao et al., 2021), which contains short-length videos, and is directly transferred to the target benchmarks without task-specific fine-tuning. And prompt templates are provided in Appendix F.

Backbones and Baselines. We use Qwen3-VL-4B, Qwen3-VL-8B (Bai et al., 2025), and InternVL3.5-8B (Wang et al., 2025c) as reasoning backbones. The navigation agent Smile is initialized from a frozen Qwen3-VL-2B backbone with parameter-efficient LoRA adapters. We compare NavIAgent with uniform sampling, query-aware keyframe selection methods, and representative video agent frameworks. We fix the maximum budget to eight frames per candidate segment such that thirty-two frames corresponds to four segments of eight frames with other frames configured similarly. Detailed implementation and hardware settings are provided in Appendix D.1.

Table 1: Comparison of accuracy with open-source video models, keyframe selection, and agent frameworks on long video benchmarks. Methods with dynamic frame allocation report their empirical average frames (*,†).

Framework / Method	Size	Frames	LongVideo Bench	LVBench	Video-MME	
					Long	Medium
Standard Video-LLMs						
VideoChat-R1 (Li et al., 2025)	7B	32	49.1	34.3	46.2	–
Video-R1 (Feng et al., 2026)	7B	32	52.7	35.3	48.2	–
Kangaroo (Liu et al., 2026a)	8B	64	54.8	39.4	49.3	55.3
Passive Agent Frameworks						
VideoMind (Liu et al., 2026b)	7B	32	56.3	40.8	49.2	–
MACF (Chen et al., 2026b)	8B	16	56.8	40.2	–	–
Graph-based Agent						
EGAgent (Rege et al., 2026)	8B	50	–	–	56.4	–
Active Video Perception						
FrameThinker (He et al., 2026b)	7B	23†	52.9	36.6	47.6	–
VideoBrain (Zou et al., 2026a)	8B	21†	53.3	41.3	49.8	–
LongVideo-R1 (Qiu et al., 2026)	7B	32	–	43.0	57.1	–
Qwen3-VL-4B Series						
Baseline	4B	32	52.5	41.3	50.7	56.6
NavIAgent	**4B**	**24.8***	**61.2** (8.7↑)	**50.7** (9.4↑)	**59.7** (9.0↑)	**67.7** (11.1↑)
Qwen3-VL-8B Series						
Baseline	8B	32	56.3	43.8	53.2	59.5
QCA (Peng et al., 2026)	8B	64	**66.9**	51.8	59.9	–
NavIAgent	**8B**	**25.2***	65.4 (9.1↑)	**53.6** (9.8↑)	**60.9** (7.7↑)	**71.1** (11.6↑)
InternVL3.5-8B Series						
Baseline	8B	64	62.7	42.8	54.3	61.3
QCA	8B	64	63.5	50.0	57.4	–
NavIAgent	**8B**	**28.4***	**63.7** (1.0↑)	**51.3** (8.5↑)	**57.6** (3.3↑)	**68.3**(7.0↑)

*The value represents the average frames under a maximum exploration budget cap of $T = 32$.
†Indicates an approximate frame count.

4.2 Main Results

Overall performance. Table 1 shows that NavIAgent consistently improves over uniform sampling across backbones without increasing the frame budget. For Qwen3-VL-4B, NavIAgent improves accuracy in Video-MME Long from 50.7% to 59.7% and LVBench from 41.3% to 50.7%. With Qwen3-VL-8B, it achieves 60.9% on Video-MME Long using only 25.2 frames, outperforming QCA at 59.9% with 64 frames. The gains are smaller for InternVL3.5-8B, but NavIAgent still improves consistently across the evaluated benchmarks, demonstrating that the proposed coarse-to-fine retrieval and active navigation pipeline is effective across different multimodal backbones. And we provide a case study in Appendix E.3.

4.3 Inference Latency Analysis

Runtime and accuracy tradeoff. Table 3 compares the inference time, frame consumption, and accuracy on Video-MME Long. Faster methods like VideoBrain require less time, but reach only 49.8% accuracy compared to 60.9% from NavIAgent. At the same time, slower methods such as LongVideo-R1 and EGAgent use more frames and take longer to run, but still achieve lower accuracy than our method. Overall, NavIAgent achieves a well-balanced tradeoff between inference speed and reasoning accuracy.



Figure 4: Performance under different frame budgets on Video-MME Long without subtitles.

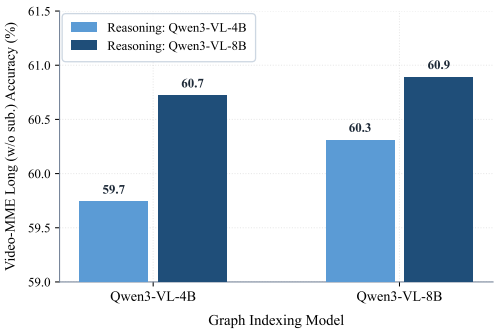

Figure 5: Interaction between graph and reasoning model scales on Video-MME Long.

Table 2: Ablation of active navigation on Video-MME Long.

Backbone	Smile	LVBench	Video-MME
Qwen3-VL-4B	✗	44.6	54.1
	✓	**50.7**	**59.7**
Qwen3-VL-8B	✗	48.2	55.3
	✓	**53.6**	**60.9**

Table 3: Inference latency and accuracy on Video-MME Long. **Bold** and underline indicate the best and second-best results.

Method	Frames	Runtime (s)	Acc. (%)
VideoBrain	**21**	**51.3**	49.8
LongVideo-R1	32	117	57.1
EGAgent	50	149	56.4
NavIAgent	25.2	111	**60.9**

4.4 Frame Budget and Model Scale Synergy

Scaling with visual budgets. Figure 4 shows that NavIAgent achieves 56.7% accuracy with only 16 frames budget, exceeding the 64-frame uniform baseline of 55.3%. Accuracy increases to 60.9%, 62.4%, and 62.6% with 32, 64, and 128 frames, showing that active navigation concentrates limited visual budgets on informative evidence and exhibits diminishing returns at larger budgets.

Interaction between retrieval and reasoning capacity. Figure 5 shows that scaling either the graph indexing model or reasoning backbone improves performance. The Qwen3-VL-4B indexing and Qwen3-VL-4B reasoning configuration already achieves 59.7%, while jointly scaling both components to Qwen3-VL-8B reaches 60.9%. This indicates that the coarse-to-fine pipeline remains effective even when powered by lightweight multimodal backbones.

4.5 Ablation on Active Navigation

Effect of closed-loop navigation. Table 2 isolates the contribution of Smile while keeping the graph retrieval stage unchanged. Specifically, activating Smile yields improvements of 6.1% on LVBench and 5.6% on Video-MME for Qwen3-VL-4B, with achieves gains of 5.4% on LVBench and 5.6% on Video-MME for Qwen3-VL-8B. These consistent gains demonstrate that closed-loop active sensing effectively guides temporal transitions to isolate answer-critical evidence. The ablation experiments only using Smile are shown in Appendix E.

5 Conclusion

In this paper, we address the core challenge in long-form video understanding where active perception must search large temporal spaces while precisely capturing short answer-critical events. To address this tradeoff, we propose NavIAgent, a coarse-to-fine active frame selection framework supported by theoretical rationale. NavIAgent decouples evidence localization into structured multimodal event graph retrieval and localized closed-loop sensing via the Smile agent, which is optimized through dual-scale reinforcement learning with epistemic intrinsic rewards. This design enables efficient long-range search and fine-grained evidence acquisition within bounded candidate segments. Extensive evaluations on Video-MME, LongVideoBench, and LVBench show that NavIAgent improves the accuracy of base vision-language models while substantially reducing visual frame consumption and inference latency. A detailed discussion of the limitations of our framework and future research directions is provided in Appendix A.

AI USAGE STATEMENT

We used AI-based language tools solely for proofreading and polishing the manuscript, including improving grammar, clarity, and overall readability. All research ideas, methodology, experiments, analyses, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of this paper.

REFERENCES

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

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29071–29082. IEEE, 2025.

Wang Chen, Yongdong Luo, Yuhui Zeng, Luojun Lin, Tianyu Xie, Fei Chao, Rongrong Ji, and Xiawu Zheng. Event-anchored frame selection for effective long-video understanding. *arXiv preprint arXiv:2603.00983*, 2026a.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870, 2026b.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11 (2):127–138, 2010.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18959–18969. IEEE, 2025.

Haichen He, Jiayi Zhou, Sifeng Shang, Yihan Hu, Yuanhan Zhang, and Kaiyang Zhou. Videoodyssey: A benchmark for ultra-long-context and omni-modal video understanding. *arXiv preprint arXiv:2605.22907*, 2026a.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. In *International Conference on Learning Representations*, volume 2026, pp. 74904–74933, 2026b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jitesh Jain, Jialuo Li, Zixian Ma, Jieyu Zhang, Chris Dongjoo Kim, Sangho Lee, Rohun Tripathi, Tanmay Gupta, Christopher Clark, Humphrey Shi, et al. Sage: Training smart any-horizon agents for long video reasoning with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41478–41488, 2026.

Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, et al. Storm: Token-efficient long video understanding for multimodal llms. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 5889–5900. IEEE, 2025.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input: J. liu et al. *International Journal of Computer Vision*, 134(3):114, 2026a.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327. IEEE, 2025.

Ye Liu, Kevin Qinghong Lin, Chang-Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 57481–57506, 2026b.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946. IEEE, 2025.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *Advances in Neural Information Processing Systems*, 38:138605–138632, 2026.

Jun Peng, Baiyang Song, Jie Li, Hui Li, Yiyi Zhou, Rongrong Ji, and Yonghong Tian. Qca: Query-and content-aware keyframe selection for long video understanding. *arXiv preprint arXiv:2607.00983*, 2026.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. Longvideo-r1: Smart navigation for low-cost long video understanding. *arXiv preprint arXiv:2602.20913*, 2026.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46575–46602, 2026.

Wenhui Tan, Ruihua Song, Jiaze Li, Jianzhong Ju, and Zhenbo Luo. Think-clip-sample: Slow-fast frame selection for video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12157–12161. IEEE, 2026.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. Allvb: All-in-one long video understanding benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7211–7219, 2025.

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

Zhou Tao, Shida Wang, Yongxiang Hua, Haoyu Cao, and Linli Xu. Dig: Differential grounding for enhancing fine-grained perception in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1695–1705, 2026.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in neural information processing systems*, 38:4350–4376, 2026a.

Shaoguang Wang, Weiyu Guo, Ziyang Chen, Xuming Hu, and Hui Xiong. Where to focus: Query-modulated multimodal keyframe selection for long video understanding. *arXiv preprint arXiv:2604.17422*, 2026b.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025a.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025c. URL https://arxiv.org/abs/2508.18265.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9088–9099, 2026c.

Fan Wei, Siru Zhong, Runmin Dong, Miao Yang, Zhaoyang Luo, and Haohuan Fu. Ground, cover, and refine: Evidence-centric frame selection for long-video question answering. *arXiv preprint arXiv:2608.01660*, 2026.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, et al. Native active perception as reasoning for omni-modal understanding. *arXiv preprint arXiv:2606.19341*, 2026.

Yifeng Yao, Yike Yun, Jing Wang, Huishuai Zhang, Dongyan Zhao, Ke Tian, Zhihao Wang, Minghui Qiu, and Tao Wang. K-frames: Scene-driven any-k keyframe selection for long video understanding. *arXiv preprint arXiv:2510.13891*, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591. IEEE, 2025.

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

Tatiana Zemskova, Solomon Andryushenko, Ilya Obrubov, Viktoriia Khoruzhaia, Ekaterina Eroshenko, Ekaterina Derevyanka, and Dmitry Yudin. Focusgraph: Graph-structured frame selection for embodied long video question answering. *arXiv preprint arXiv:2603.04349*, 2026.

Ce Zhang, Jinxi He, Katia Sycara, and Yaqi Xie. Lens: Adaptive spatio-temporal zooming for keyframe sampling in long-form videos. *arXiv preprint arXiv:2607.25125*, 2026a.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. In *International Conference on Learning Representations*, volume 2026, pp. 140958–140985, 2026b.

Junyang Zhang, Puhan Luo, Chen Tang, Yuxi Shi, and Xiang-Yang Li. Coverage-driven adaptive keyframe selection for video understanding. *arXiv preprint arXiv:2608.00714*, 2026c.

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.

Junbo Zou, Ziheng Huang, Shengjie Zhang, Liwen Zhang, and Weining Shen. Videobrain: Learning adaptive frame sampling for long video understanding. *arXiv preprint arXiv:2602.04094*, 2026a.

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

A Limitations and Future Work

The downstream reasoning accuracy of NavIAgent is bounded by the retrieval precision and recall of the coarse stage multimodal event graph, as the fine grained active exploration agent operates strictly within the localized candidate segments. Furthermore, the topological quality of the event graph, particularly the establishment of causal dependencies and entity coreference relations across video chunks, depends substantially on the perceptual and reasoning capabilities of the underlying base vision language model. In future work, we plan to explore algorithmic enhancements to boost graph retrieval robustness, including learnable graph traversal policies optimized via reinforcement learning, uncertainty aware topological representations to tolerate foundation model noise, and bidirectional feedback mechanisms that enable the active agent to re-examine alternative intervals when localized evidence remains insufficient.

B Theoretical Analysis and Proofs

B.1 Proof of Proposition 1: Expressive Separation

Formal Setting. Let a video be an indexed sequence of frames $V = (f_1, \dots, f_N)$, $Q \in \mathcal{Q}$ be the question, and $y \in \mathcal{Y}$ be the ground-truth answer with prior uncertainty

$$\varepsilon_{\mathcal{Y}} := 1 - \sup_{y \in \mathcal{Y}} P(y) > 0.$$

Querying frame index $i_t \in 1, \dots, N$ yields a distilled observation $o_{i_t} = \mathcal{Z}(f_{i_t}) \in \mathcal{O}$. For an exploration budget $T \geq 2$, a passive policy $\pi^{\text{pass}} = (\psi_t{}_{t=1}^{T}, h) \in \Pi^{\text{pass}}(T)$ commits all T indices in advance as $\psi_t(Q)$ before observing any frames, predicting

$$\hat{y} = h(Q, o_{i_1}, \dots, o_{i_T}).$$

An active policy $\pi^{\text{act}} = (\phi_1, \phi_t{}_{t=2}^{T}, h) \in \Pi^{\text{act}}(T)$ selects indices sequentially via

$$i_1 = \phi_1(Q), \qquad i_t = \phi_t(Q, o_{i_1}, \dots, o_{i_{t-1}}), \quad t \geq 2.$$

The population risk over instance distribution $\mathcal{D}$ is

$$\mathcal{L}(\pi; \mathcal{D}) = \mathbb{E}_{(V,Q,y)\sim\mathcal{D}} \left[\mathbf{1}(\pi(V, Q) \neq y) \right].$$

Proof. We first establish the inclusion $\Pi^{\text{pass}}(T) \subseteq \Pi^{\text{act}}(T)$. For any passive policy $\pi^{\text{pass}} \in \Pi^{\text{pass}}(T)$ parameterized by functions $\psi_t{}_{t=1}^{T}$, define an active policy $\tilde{\pi}^{\text{act}}$ with

$$\tilde{\phi}_1(Q) = \psi_1(Q), \qquad \tilde{\phi}_t(Q, o_{i_1}, \dots, o_{i_{t-1}}) = \psi_t(Q)$$

for all $t \geq 2$. Thus, $\tilde{\pi}^{\text{act}}$ ignores the observation history and produces the same query sequence and prediction as π^{pass} on every instance. Hence,

$$\Pi^{\text{pass}}(T) \subseteq \Pi^{\text{act}}(T).$$

To establish strictness, consider a distribution $\mathcal{D} * \text{sep}$ containing an initial clue frame $f * \text{clue}$ and $T + 1$ candidate frames $\mathcal{C} = c_1, \dots, c_{T+1}$. Let the target index

$$k^* \sim \text{Uniform}(\{1, \dots, T + 1\})$$

be independent of Q, and let y be sampled independently from $P_{\mathcal{Y}}$. The clue observation reveals the target index,

$$o_{\text{clue}} = k^*,$$

while the candidate observations satisfy

$$o_{c_{k^*}} = y, \qquad o_{c_k} = \varnothing, \quad k \neq k^*,$$

where $\varnothing$ is uninformative noise independent of y.

An active policy can first query f_{clue} and then select c_{k^*} based on the observed clue. Since $T \geq 2$, it can acquire y within the exploration budget and therefore achieve

$$\inf_{\pi \in \Pi^{\text{act}}(T)} \mathcal{L}(\pi; \mathcal{D}_{\text{sep}}) = 0.$$

In contrast, a passive policy must determine its complete query set $S = i_1, \ldots, i_T$ from Q alone. Since k^* is uniform and independent of Q, at most T of the $T+1$ candidate frames can be covered, giving

$$\mathbb{P}(c_{k^*} \notin S) \geq 1 - \frac{T}{T+1} = \frac{1}{T+1}.$$

Whenever $c_{k^*} \notin S$, all observed candidate frames are independent of y, so the optimal prediction error is at least the prior uncertainty $\varepsilon_{\mathcal{Y}}$. Therefore,

$$\mathcal{L}(\pi^{\text{pass}}; \mathcal{D}_{\text{sep}}) \geq \frac{\varepsilon_{\mathcal{Y}}}{T+1} > 0.$$

Hence,

$$\inf_{\pi \in \Pi^{\text{pass}}(T)} \mathcal{L}(\pi; \mathcal{D}_{\text{sep}}) > \inf_{\pi \in \Pi^{\text{act}}(T)} \mathcal{L}(\pi; \mathcal{D}_{\text{sep}}),$$

which proves

$$\Pi^{\text{pass}}(T) \subsetneq \Pi^{\text{act}}(T).$$

□

B.2 Proof of Proposition 2: Resolution Dilution

Proof. Let the continuous video timeline be represented by $[0, L]$. An answer-critical event occupies the interval

$$\mathcal{I}^* = [\tau^*, \tau^* + w_{\text{event}}],$$

where $w_{\text{event}} > 0$ and

$$\tau^* \sim \mathcal{U}(0, L - w_{\text{event}}).$$

Consider an agent executing a closed-loop active policy with at most T point inspections at temporal locations

$$\mathcal{X} = \{x_1, \ldots, x_T\} \subseteq [0, L],$$

where each inspection x_t is adaptively chosen based on historical observations $\mathcal{F}_{t-1} = (x_1, o_1, \ldots, x_{t-1}, o_{t-1})$. By assumption, background observations outside $\mathcal{I}^*$ provide zero mutual information about τ^* ($I(o_t; \tau^*) = 0$).

An inspection at x_t intercepts the event whenever

$$x_t \in [\tau^*, \tau^* + w_{\text{event}}],$$

or equivalently,

$$\tau^* \in [x_t - w_{\text{event}}, x_t].$$

Thus, each inspection covers at most a temporal region of measure w_{event} in the space of possible event locations.

The set of event locations that can be intercepted by the complete trajectory is contained in

$$\mathcal{S} = \bigcup_{t=1}^{T} [x_t - w_{\text{event}}, x_t].$$

By subadditivity of the Lebesgue measure, regardless of the adaptive choices of x_t,

$$\mu(\mathcal{S}) \leq T w_{\text{event}}.$$

Under the uninformative background assumption, failing to intercept the event up to step t reveals only that $\tau^* \notin \bigcup_{k=1}^{t} [x_k - w_{\text{event}}, x_k]$, leaving the posterior distribution of τ^* uniform over the remaining domain. Therefore, the overall interception probability satisfies

$$P_{\text{hit}} \leq \frac{\mu(\mathcal{S})}{L - w_{\text{event}}} \leq \frac{T w_{\text{event}}}{L - w_{\text{event}}}.$$

Therefore,

$$P_{\text{miss}} = 1 - P_{\text{hit}} \geq 1 - \frac{T w_{\text{event}}}{L - w_{\text{event}}}.$$

For fixed T and w_{event}, taking $L \to \infty$ gives

$$\lim_{L \to \infty} P_{\text{miss}} = 1.$$

Now consider percentage-based navigation with discrete action magnitudes $a_t \in \mathcal{A}$ and

$$\delta_{\min} = \min_{a \in \mathcal{A}} |a| > 0.$$

The corresponding physical displacement satisfies

$$|\Delta \tau_t| = |a_t| L \geq \delta_{\min} L.$$

Hence, when

$$L > \frac{w_{\text{event}}}{\delta_{\min}},$$

the minimum physical displacement exceeds the duration of the target event:

$$|\Delta \tau_t| > w_{\text{event}}.$$

Algorithm 1 NavIAgent Inference Pipeline

Require: Video V, question Q, multimodal memory graph $\mathcal{G} = (\mathcal{C}, E)$ with chunk records $\{m_i\}_{i=1}^{M}$, aligned speech transcripts $\mathcal{U} = \{\mathcal{U}_i\}_{i=1}^{M}$, trained navigation policy π_θ
Ensure: Generated answer y

$\eta \leftarrow \mathrm{Analyze}(Q, \mathcal{G})$ ▷ Question classification and coreference resolution
if $\eta.\mathrm{type} = \texttt{Global}$ **then**
  return $\mathrm{LLM}(Q, \mathrm{FormatTimeline}(\mathcal{G}))$ ▷ Direct global reasoning over memory graph
end if
$\{(\mathrm{task}_j, \Omega_j)\}_{j=1}^{J} \leftarrow \mathrm{Planner}(Q, \eta)$ ▷ Query decomposition into retrieval subtasks with tool set Ω_j
$\mathcal{W}_{\mathrm{ret}} \leftarrow \emptyset$
$\mathcal{C}_Q \leftarrow \emptyset$
for $j = 1$ **to** J **do**
  for tool $\tau \in \Omega_j$ **in parallel do**
    $(\mathrm{ev}_\tau, \mathcal{C}_\tau) \leftarrow \mathcal{R}_\tau(\mathrm{task}_j, \mathcal{G}, \mathcal{U})$
    $\mathcal{W}_{\mathrm{ret}} \leftarrow \mathcal{W}_{\mathrm{ret}} \cup \mathrm{ev}_\tau$
    $\mathcal{C}_Q \leftarrow \mathcal{C}_Q \cup \mathrm{Expand}_{\mathcal{G}}(\mathcal{C}_\tau)$
  end for
  if $\mathrm{Grader}(Q, \mathcal{W}_{\mathrm{ret}}) = \texttt{Sufficient}$ **then** ▷ Early exit if retrieval is sufficient
    break
  end if
end for
for each candidate chunk $c_i \in \mathcal{C}_Q$ **in parallel do** ▷ Local fine-grained active navigation
  $(\mathcal{K}_i, \{d_k\}_{k \in \mathcal{K}_i}) \leftarrow \mathrm{Smile}(\pi_\theta,\, c_i,\, Q)$
end for
$\mathcal{W} \leftarrow \mathrm{Concat}_{c_i \in \mathcal{C}_Q} \left(m_i \circ [d_k]_{k \in \mathcal{K}_i} \circ [u]_{u \in \mathcal{U}_i} \right)$ ▷ Multimodal context aggregation in Eq. 15
return $\mathrm{VLM}(Q, \mathcal{K}, \mathcal{W})$ ▷ Conditioned final answer generation

Between two consecutive jump points x_t and x_{t+1}, this induces non-interceptible blind intervals of measure at least $(\delta_{\min} L - w_{\mathrm{event}})$, bounding the single-stride interception probability by $\frac{w_{\mathrm{event}}}{\delta_{\min} L} \to 0$. Thus, the physical temporal resolution of the fixed percentage action space becomes coarser than the target event as L increases, establishing the claimed resolution dilution. □

C ALGORITHM

C.1 NavIAgent Inference Algorithm

Algorithm 1 summarizes the overall inference pipeline of NavIAgent. Given a precomputed temporal memory graph $\mathcal{G}$ and a user question Q, the planner first decomposes the query to retrieve and expand candidate video chunks $\mathcal{C}_Q$. Smile is subsequently dispatched in parallel across these candidate segments to actively navigate and select answer-diagnostic keyframes while generating concise visual descriptions. Finally, the aggregated multimodal context $\mathcal{W}$, which consolidates chunk summaries, navigated keyframe descriptions, and time-aligned speech transcripts, is provided to the VLM to produce the final answer.

C.2 Smile Training Algorithm

Algorithm 2 details the dual-scale GRPO training procedure of Smile policy π_θ. For each training instance, K parallel trajectories are rolled out with active transitions evaluated via step-level rewards balancing visual diversity, belief updates, and jump costs. Trajectories terminate with semantic verification rewards. A dual-scale advantage decouples group-level global success from step-level navigation efficiency, driving policy updates via clipped surrogate optimization with KL constraints against reference model π_{ref}.

Algorithm 2 Smile Dual-Scale GRPO Training

Require: Navigation policy π_θ with LoRA, reference policy π_{ref}, training set $\mathcal{D}$, group size K, learning rate η, clipping threshold ϵ_{clip}, discount factor γ, reward weights $(\beta, \alpha, \lambda, \lambda_{\text{term}})$, advantage balance ω, KL coefficient β_{KL}, entropy coefficient β_{ent}, belief momentum ρ, max exploration budget T

Ensure: Optimized policy π_θ

repeat
  Sample batch $\{(V_n, Q_n, y_n^*)\}$ from $\mathcal{D}$
  for each sample (V_n, Q_n, y_n^*) in batch **do**
    for rollout index $k = 1$ **to** K **do**
      Initialize memory $\mathcal{M}_0 \leftarrow \{Q_n\}$
      Initialize frame $f_{n,k,1}$ and belief $b_0 \leftarrow 0$
      for step $t = 1$ **to** T **do**
        Sample action $a_{n,k,t} \sim \pi_\theta^{\text{action}}(\cdot \mid \mathcal{M}_{t-1}, f_{n,k,t})$
        Generate summary $d_{n,k,t} \sim \pi_\theta^{\text{desc}}(\cdot \mid \mathcal{M}_{t-1}, f_{n,k,t})$
        Update memory $\mathcal{M}_t \leftarrow \mathcal{M}_{t-1} \cup \{d_{n,k,t}\}$
        Compute belief b_t using Eq. 10 with momentum ρ
        if $a_{n,k,t} = \texttt{STOP}$ **then**
          $R_{n,k,t} \leftarrow \alpha \cdot \max(b_t - b_{t-1}, 0)$ ▷ Terminal step with belief gain only
          $T_{n,k} \leftarrow t$
          break
        else
          $f_{n,k,t+1} \leftarrow \text{Env. step}(a_{n,k,t})$ ▷ Advance video progress bar
          $R_{n,k,t} \leftarrow \beta \cdot \|\text{CLIP}_{\text{img}}(f_{n,k,t+1}) - \text{CLIP}_{\text{img}}(f_{n,k,t})\|_2 + \alpha \cdot \max(b_t - b_{t-1}, 0) - \lambda \cdot |a_{n,k,t}|$
          if $t = T$ **then**
            $T_{n,k} \leftarrow T$
          end if
        end if
      end for
      Compute terminal semantic verification reward $R_{\text{term},n,k} \in [0, 1]$ via ground truth y_n^*
      $G_{n,k} \leftarrow \sum_{t=1}^{T_{n,k}} \gamma^{t-1} R_{n,k,t} + \lambda_{\text{term}} \cdot R_{\text{term},n,k}$ ▷ Total discounted return
    end for
    Compute group return statistics $\mu_G \leftarrow \frac{1}{K} \sum_{k=1}^{K} G_{n,k}$ and $\sigma_G \leftarrow \text{std}(\{G_{n,k}\}_{k=1}^{K})$
    Compute step reward statistics across active transitions μ_{step} and σ_{step}
    for $k = 1$ **to** K **do**
      $A_{n,k}^{\text{macro}} \leftarrow \frac{G_{n,k} - \mu_G}{\sigma_G}$ ▷ Macro trajectory advantage
      for $t = 1$ **to** $T_{n,k}$ **do**
        $A_{n,k,t}^{\text{micro}} \leftarrow \frac{R_{n,k,t} - \mu_{\text{step}}}{\sigma_{\text{step}}}$ ▷ Micro step advantage
        $A_{n,k,t} \leftarrow A_{n,k}^{\text{macro}} + \omega \cdot A_{n,k,t}^{\text{micro}}$ ▷ Dual-scale advantage in Eq. 12
        $r_{n,k,t}(\theta) \leftarrow \frac{\pi_\theta(a_{n,k,t}, d_{n,k,t} \mid \mathcal{M}_{t-1}, f_{n,k,t})}{\pi_{\theta_{\text{old}}}(a_{n,k,t}, d_{n,k,t} \mid \mathcal{M}_{t-1}, f_{n,k,t})}$
      end for
    end for
    Compute GRPO objective $\mathcal{J}(\theta)$ over group rollouts via Eq. 13 with $\epsilon_{\text{clip}}, \beta_{\text{KL}}, \beta_{\text{ent}}, \pi_{\text{ref}}$
    $\theta \leftarrow \theta + \eta \cdot \nabla_\theta \mathcal{J}(\theta)$ ▷ Backpropagation with gradient clipping
  end for
until converged

D Detailed Implementation and Training Setup

D.1 Hardware and Hyperparameter Settings

All training and inference experiments are conducted on a single computation node equipped with 8 Ascend 910B3 NPUs (64 GB VRAM per card). The Smile active exploration policy π_θ is initialized

Table 4: Smile reinforcement learning training hyperparameters.

Category	Hyperparameter	Value
LoRA & Optimizer	LoRA rank r	32
	LoRA alpha	64
	Policy learning rate η	1×10^{-5}
	Gradient clipping threshold $c_{\max}$	1.0
Rollout & GRPO	Batch size B	8
	Group rollout size K	8
	Max exploration budget T	50
	GRPO clipping threshold ϵ_{clip}	0.2
	Micro advantage balance ω	0.5
	KL divergence coefficient β_{KL}	0.01
	Entropy bonus coefficient β_{ent}	0.01
	Discount factor γ	0.9
Reward Formulation	Visual novelty weight β	0.5
	Information gain weight α	2.0
	Terminal verification weight λ_{term}	4.0
	Belief momentum coefficient ρ	0.8
	Belief ratio clipping cap C	2.0
	Numerical stabilizer ε_{eps}	1×10^{-6}
Generation Sampling	Base temperature	1.0
	Temperature ramp coefficient	0.2

Table 5: NavIAgent inference and execution hyperparameters.

Hyperparameter	Value
Input frame resolution $H \times W$	288×288
Graph expansion context hops	1
Candidate chunk set max limit $\lvert\mathcal{C}_Q\rvert$	4
Max active navigation steps per chunk (T)	8
Discrete action space $\mathcal{A}$	$\{+3\%, +5\%, +8\%, +12\%, +20\%, \texttt{STOP}\}$
Shot boundary threshold θ_{cut}	35.0
seg-min-scene-duration	40
Final aggregation max frames per VLM call	32

from Qwen3-VL-2B and fine-tuned using LoRA (Hu et al., 2021) applied to the attention projection matrices. The fine-grained policy is integrated into the coarse-to-fine NavIAgent pipeline during inference. We detail the reinforcement learning training hyperparameters of Smile in Table 4 and the downstream inference configurations of NavIAgent in Table 5. According to the definition of action space $\mathcal{A}$, it is notable that candidate chunks average only around one minute, so the smallest 3% step equals less than two seconds, which is fine enough to catch brief moments and prevents the model from getting stuck in back-and-forth loops.

D.2 Belief State Estimation and Distributed Reward Calculation

In this subsection, we describe the exact operational computation of the sufficiency metrics p_1 and p_2 used for the recursive belief update in Eq. 10.

Smooth sufficiency scoring. We employ a frozen Qwen3-VL-8B-Instruct model $\mathcal{J}$ as a calibrated diagnostic verifier, served via a vLLM backend with `top_logprobs=5`. Given the query context Q, the accumulated textual memory $\mathcal{M}_t$, and optionally the ground-truth target y^*, $\mathcal{J}$ is prompted to output a single categorical token indicating whether the context contains sufficient clues to resolve the question. To avoid non-differentiable or high-variance binary decisions, we extract the top-K log-probabilities $\{(\tilde{v}_k, \ell_k)\}_{k=1}^{K}$ (with $K = 5$) corresponding to the first generated token. Let $\mathcal{V}_+ = \{\texttt{"1"}\}$ denote the affirmation token set (output by $\mathcal{J}$ when the context is deemed sufficient) and $\mathcal{V}_- = \{\texttt{"0"}\}$ denote the insufficiency token set. The smooth diagnostic score $p \in [0, 1]$ is

obtained via softmax normalization over the valid token subspace:

$$p = \frac{\sum_{k:\,\tilde{v}_k \in \mathcal{V}_+} \exp(\ell_k)}{\sum_{k:\,\tilde{v}_k \in \mathcal{V}_+} \exp(\ell_k) \;+\; \sum_{k:\,\tilde{v}_k \in \mathcal{V}_-} \exp(\ell_k)}. \tag{16}$$

If the denominator collapses to zero (i.e., neither affirmation nor insufficiency appears in the top-K candidates), p defaults to 0.5 as an uninformative prior.

Counterfactual diagnostic probing. At transition step t, the belief tracker issues two queries against the verifier $\mathcal{J}$:

$$p_1 = \mathcal{J}(Q, \mathcal{M}_t), \quad p_2 = \mathcal{J}(Q, \mathcal{M}_t, y^*), \tag{17}$$

where p_1 measures whether the accumulated memory $\mathcal{M}_t$ provides sufficient standalone context to resolve Q open-endedly, and p_2 evaluates whether $\mathcal{M}_t$ decisively corroborates the specific ground-truth answer y^*. The ratio $p_2/\max(p_1, \varepsilon_{\text{eps}})$ (with $\varepsilon_{\text{eps}} = 10^{-6}$) isolates task-relevant evidentiary diagnosticity:

- When visited frames accumulate decisive evidence directly supporting y^*, the verification certainty spikes, driving p_2 to high values while narrowing the open-ended ambiguity, thereby elevating the ratio p_2/p_1.
- When visited frames are irrelevant or contain misleading distractors, the context fails to corroborate y^* ($p_2 \to 0$), keeping the ratio depressed near zero.

The ratio is strictly bounded by threshold $C = 2.0$ to suppress outlier belief updates, driving the moving-average belief update formalized in Eq. 10.

D.3 Temporal Segmentation and Memory Graph Construction

The structured chunk set $\mathcal{C} = \{c_1, \ldots, c_M\}$ introduced in Sec. 3.2 is partitioned through a dual-stage procedure comprising content-based visual shot detection and graph topological instantiation.

Content-based shot detection. We employ PySceneDetect equipped with the HSV ContentDetector to establish atomic chunk boundaries. For adjacent frames f_{t-1} and f_t transformed into HSV color space, the detector calculates the mean frame difference:

$$\delta_t = \frac{1}{3} \sum_{c \in \{H,S,V\}} \frac{1}{|\Omega|} \sum_{p \in \Omega} \left| f_t^{(c)}(p) - f_{t-1}^{(c)}(p) \right|, \tag{18}$$

where Ω denotes the pixel lattice. A frame index t is registered as an accepted boundary if its visual discrepancy exceeds threshold θ_{cut} while satisfying a minimum shot duration constraint $L_{\min}$:

$$\mathcal{B} = \left\{ t \;\middle|\; \delta_t > \theta_{\text{cut}},\; t - t^{\mathcal{B}}_{\text{prev}} \geq L_{\min} \right\}, \tag{19}$$

where $t^{\mathcal{B}}_{\text{prev}}$ denotes the immediate preceding boundary. The resulting ordered intervals $[t_i^{\text{start}}, t_i^{\text{end}}]$ define the temporal boundaries of atomic chunk $c_i \in \mathcal{C}$. If $\mathcal{B} = \emptyset$, the entire video horizon is treated as a single unified chunk.

Multimodal graph topological wiring. Each chunk c_i is mapped to a memory vertex $m_i = \langle \sigma_i, [t_i^{\text{start}}, t_i^{\text{end}}], \ell_i, \xi_i \rangle$. The multimodal edge set $E = E_{\text{causal}} \cup E_{\text{coref}} \cup E_{\text{temporal}}$ of graph $\mathcal{G} = (\mathcal{C}, E)$ is instantiated as follows:

- **Temporal contiguity edges (E_{temporal}):** Bidirectional edges (c_i, c_{i+1}) connecting strictly consecutive segments to preserve local chronological flow.
- **Entity coreference edges (E_{coref}):** Directed edges (c_i, c_j) established when the primary subject entities satisfy $\text{IoU}(\xi_i, \xi_j) > 0.5$ or trigger an LLM-verified nominal coreference link across non-adjacent segments.
- **Causal progression edges (E_{causal}):** Directed edges (c_i, c_j) instantiated when an auxiliary LLM prompts verify that an event in c_i is a prerequisite or direct cause for the event summarized in σ_j.

During inference, candidate segment expansion $\text{Expand}_{\mathcal{G}}(\mathcal{C}_\tau)$ queries 1-hop incident edges across E, ensuring that fine-grained active navigation inherits local causal and entity context without breaking temporal continuity.

Table 6: Validation downstream accuracy (%) across training steps (mean ± standard deviation over 3 random seeds).

Step	Accuracy (%)	Step	Accuracy (%)
30	32.2 ± 0.9	240	47.0 ± 0.7
60	33.5 ± 0.3	270	50.2 ± 1.3
90	34.8 ± 0.7	300	53.0 ± 1.6
120	36.2 ± 0.5	330	55.5 ± 0.5
150	37.8 ± 1.2	360	57.5 ± 0.8
180	40.1 ± 1.5	390	58.9 ± 1.4
210	43.5 ± 0.3	–	–

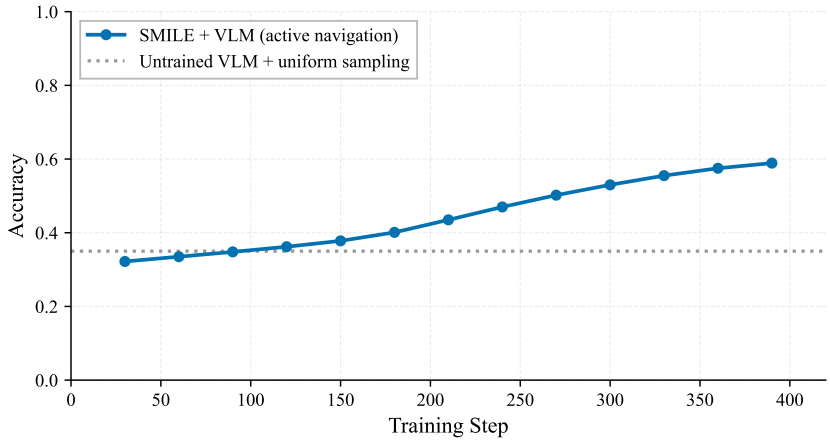

Figure 6: **Evaluation accuracy curve during RL training on NExT-QA.** Active navigation with Smile (blue) improves steadily from 32.2% to 58.9%, substantially surpassing the frozen VLM baseline (grey dashed, 35.0%).

E Additional Experimental Results and Extended Analyses

E.1 Smile Training Dynamics and Accuracy Convergence

Monotonic accuracy progression. We train Smile on the NExT-QA training set using a batch size of $B = 8$ videos with $K = 8$ parallel rollout trajectories per sample, evaluating performance on a held-out validation subset of video-question pairs. To account for training stochasticity, experiments are repeated across three independent random seeds, with the step-wise mean and standard deviation summarized in Table 6. As illustrated in Figure 6, the downstream answer accuracy of Smile rises monotonically from an initial 32.2% to 58.9%, while the frozen zero-shot VLM baseline stagnates at 35.0%. The upward trend suggests that step-level rewards provide useful training signals in this setting. Guided by the composite dense supervisory signals, the policy rapidly learns to escape uninformative background sequences and actively maneuver toward answer-diagnostic keyframes, confirming the efficacy of closed-loop exploration over static inference.

Dissecting reward signal contributions. Table 7 presents the empirical contribution breakdown of each component within the trajectory-level cumulative return $G_i = \sum_{t=1}^{T_i} \gamma^{t-1} R_{i,t} + \lambda_{\text{term}} R_{\text{term},i}$. The terminal semantic verification reward accounts for 40.9% of the global return (1.881 out of 4.600), whereas the intermediate step-level signals provide the remaining 59.1%, split between CLIP visual novelty (31.4%, contribution 1.444) and evidentiary belief gain (27.7%, contribution 1.275). While terminal correctness acts as the indispensable anchor for final task success, the two dense rewards account for a substantial fraction of the observed trajectory return. These results suggest that the dense per-step rewards make a substantial contribution to the overall training return.

Table 7: **Trajectory-level reward contribution breakdown.** Mean per-step rewards are averaged over all active transitions assuming an empirical average of 8 active steps per rollout, weighted by $\beta = 0.5$, $\alpha = 2.0$, and $\lambda_{\text{term}} = 4.0$.

Component	Weight	Per-step Mean	Trajectory Contribution	Share
CLIP Novelty	$\beta = 0.5$	0.361	$8 \times 0.5 \times 0.361 = 1.444$	31.4%
Belief Gain	$\alpha = 2.0$	0.080	$8 \times 2.0 \times 0.080 = 1.275$	27.7%
Terminal Reward	$\lambda_{\text{term}} = 4.0$	0.470	$4.0 \times 0.470 = 1.881$	40.9%
Total Positive Signals	—	—	**4.600**	**100.0%**

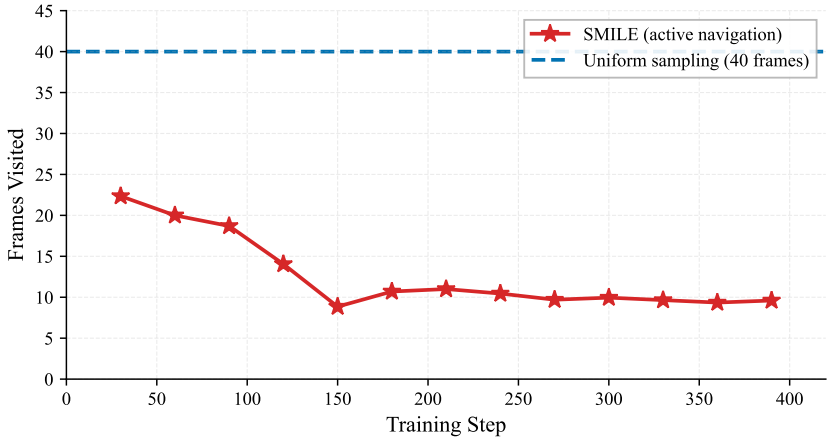

Figure 7: **Evolution of frame consumption during policy training.** Inspected frames per query (red stars) monotonically decrease from 22.4 to 9.6, achieving a 76% computational reduction compared to fixed uniform sampling (blue dashed, 40 frames).

E.2 Frame Efficiency Improves Alongside Accuracy

Simultaneous efficiency gain and evidence localization. Figure 7 profiles the average cumulative number of frames inspected per video question across training checkpoints within the NavI-Agent pipeline. Concurrently with the accuracy surge from 32.2% to 58.9%, the average frame footprint drops sharply from 22.4 down to 9.6 frames per query, which corresponds to a 57.1% reduction in frame inspection overhead. In contrast to the uniform sampling baseline, which rigidly consumes a fixed budget of 40 frames, our framework attains superior task accuracy while utilizing approximately 76% fewer visual inputs. This inverse trend suggests that higher accuracy does not necessarily require proportionally more temporal sampling. Instead, it indicates that Smile transitions from unfocused exploration to targeted perceptual search, acquiring the capability to eliminate temporal redundancy and rapidly home in on answer-bearing events.

Equilibrium between exploration incentive and economy. Throughout the later stages of policy optimization, the average inspected frame count stabilizes around 9.6 frames rather than continuously decreasing toward very short trajectories, such as one or two steps. This convergence behavior demonstrates that our reward formulation provides a practical tradeoff between information gain and navigation cost. The step-jump penalty $\lambda \cdot |a_t|$ successfully deters excessive wandering, while the belief gain and visual novelty incentives are intended to discourage premature termination while maintaining sufficient temporal exploration.

E.3 Case Study

Figure 8 illustrates the execution flow of NavIAgent on a representative query. Given the question asking what the programmer is doing at 2:00 PM, the system first enters the question analysis phase, classifying it as a local query and extracting the key entity programmer alongside the temporal an-

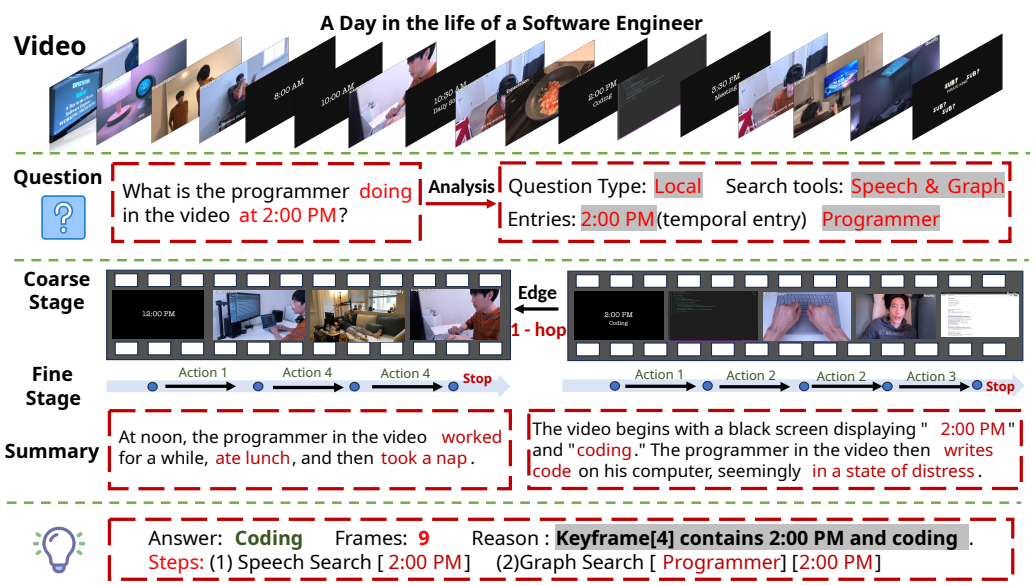

Figure 8: Qualitative case study of NavIAgent execution trajectory on a temporal grounding QA task.

chor 2:00 PM. During the coarse stage, the planner leverages speech search and multimodal event graph search to identify candidate video segments, followed by a one hop topological edge expansion to preserve causal and temporal continuity. In the subsequent fine stage, the Smile agent performs active closed loop navigation strictly within the retrieved candidate windows. By adaptively executing sequential temporal transitions and updating its textual memory, the agent efficiently locates the target moment and triggers the stop action.

Ultimately, NavIAgent isolates the decisive keyframe showing 2:00 PM and coding using only 9 inspected frames in total. The aggregated multimodal context is then supplied to the downstream vision language model to generate the correct final answer Coding with minimal frame redundancy

F Prompt Templates

F.1 Structured Memory Extraction Prompt

Prompt Template for Structured Memory Extraction

```
[{"role": "user",
  "content": You are analyzing a video clip. Extract structured information from the clip
description.

  Clip time range: {{ start_time }}s - {{ end_time }}s
  Clip description: {{ clip_description }}

  Extract the following information and return as JSON:
  - when: Time context and situation (e.g., "during the meeting", "after the event starts")
  - where: Location or setting (e.g., "in the office", "at the park", "indoor studio")
  - who_what: Main people/characters and what they are doing (e.g., "a man presenting")
  - summary: A concise summary of the key event or action in this clip (1-2 sentences)

  Return ONLY a JSON object with these 4 keys. Use empty string "" if information is not
available.

  Example output:
  {
```

```
    "when": "during the presentation",
    "where": "conference room",
    "who_what": "presenter showing slides, audience listening",
    "summary": "A presenter is delivering a presentation to an audience in a conference
room."
  }}]
```

F.2 Graph Retrieval Prompt

Prompt Template for Graph Retrieval

```
[{"role": "user",
  "content": You are a video event retrieval system. Given a user question and a list of
event summaries with their IDs, identify which events are most relevant to answering the
question.

  Question: {{ question }}

  {% if parsed_entities %}Parsed entities: {{ parsed_entities | join(", ") }}{% endif %}
  {% if parsed_time_anchor %}Parsed time anchor: {{ parsed_time_anchor }}{% endif %}
  {% if parsed_location %}Parsed location: {{ parsed_location }}{% endif %}

  === Event Summaries ===
  {% for node in node_list %}
  [{{ node.node_id }}] ({{ "%.1f"|format(node.start_time) }}s-{{
"%.1f"|format(node.end_time) }}s) {{ node.summary }}
  {% endfor %}

  === Instructions ===
  Return ONLY a JSON list of node IDs that are relevant to the question, e.g. [0, 3, 5].
  If no events are relevant, return an empty list [].
  Output ONLY the JSON list, nothing else.}]
```

F.3 SMILE Action Selection and Frame Captioning Prompts

Prompt Template for SMILE Action Selection

```
[{"role": "system",
  "content": You are an intelligent video navigation agent.},

 {"role": "user",
  "content": You are a video navigation agent. You must locate key frames to answer the
question.

  Question: <question>

  <memory_section>

  Chunk has <total_frames> frames total. You are at frame <current_frame_idx>/<total_frames>
(<current_pos>) | Step <step>/<max_steps>

  Available actions (percentage of candidate chunk):
    0 - Forward 3% (micro)
    1 - Forward 5% (small)
    2 - Forward 8% (medium)
    3 - Forward 12% (large)
    4 - Forward 20% (jump)
    5 - STOP and answer

  Output your action number after 'ACTION:'.
```

```
  Output format:
  ACTION: <number>}]
```

Prompt Template for SMILE Frame Captioning

```
[{"role": "system",
  "content": You are an intelligent video navigation agent.},
 {"role": "user",
  "content": Briefly describe what you see in this frame, focusing on details relevant to
the question: "<question>"
  Output format:
  DESCRIPTION: <one sentence>}]
```

F.4 Evidence Sufficiency Grading Prompt

Prompt Template for Evidence Sufficiency Grading

```
[{"role": "user",
  "content": You are evaluating whether the collected evidence is sufficient to answer the
question.
  Question: {{ question }}
  Collected evidence:
  {{ working_memory }}
  Determine: Can the question be answered confidently based on this evidence?
  Respond in this format:
  Line 1: "yes" (sufficient) or "no" (need more evidence)
  Line 2+: Brief reasoning
  Examples:
  yes
  The evidence clearly shows the person used a microwave at 12:15, confirmed by both graph
and visual search.
  no
  The evidence shows the person was in the kitchen, but doesn't specify which object they
used.}]
```

F.5 Final Answer Generation Prompt

Prompt Template for Final Answer Generation

```
[{"role": "user",
  "content": You are a precise video understanding assistant. Answer the question based
ONLY on the provided evidence. Do not use any external knowledge or make assumptions.
  IMPORTANT RULES:
  1. Answer ONLY based on the evidence below.
  2. If the evidence is insufficient to answer the question, respond: "Based on the
available information, I cannot determine the answer to this question."
  3. Do not speculate or infer beyond what is explicitly stated.
  4. Keep the answer concise and factual.
  === EVIDENCE ===
```

```
{{ evidence }}
=== QUESTION ===
{{ question }}
=== ANSWER ===}]
```