EVOVIDEORAG: NAVIGATE, VERIFY, AND EVOLVE FOR TOKEN-EFFICIENT VIDEO UNDERSTANDING VIA ON-POLICY VIDEO DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) grounds model generation in external evidence. With the development of multimodal models and agentic techniques, RAG has expanded from one-shot retrieval to continual evidence search, observation, and verification across heterogeneous information sources. Video is a particularly challenging modality: query-relevant information is temporally sparse and distributed across multiple scales, while each observation incurs substantial visual-token and computational costs. Existing video agents are commonly designed for individual tasks, making it difficult for a single policy to handle tasks with different input and output requirements, such as video corpus moment retrieval (VCMR) and video question answering (VideoQA). Meanwhile, successful and failed interaction trajectories are often discarded after a task, whereas continually accumulating external skills increases maintenance and invocation costs at inference time. To address these challenges, we propose EVOVIDEORAG, a unified, task-adaptive, self-evolving, and cost-aware RAG framework for video-language understanding. Specifically, (1) EVOVIDEORAG formulates heterogeneous tasks, including VCMR and VideoQA, as task-conditioned sequential evidence acquisition, enabling a shared agent to plan, retrieve, and verify video evidence and to adaptively decide whether to continue searching, adjust the temporal range, increase observation fidelity, or submit a prediction; (2) we introduce On-Policy Video Self-Distillation (OPVD), which turns new trajectories generated by the current policy into self-supervision for its successor, whose renewed interaction with the environment forms an intergenerational "interaction–distillation–reinteraction" self-evolution cycle without requiring an additional teacher or an ever-growing long-term skill library at inference time; and (3) we employ dynamic visual budgeting and cost constraints to allocate visual computation according to task difficulty and evidence sufficiency, reducing redundant frames and visual-token consumption while preserving task performance. Experiments on NExT-QA, Video-MME, Charades-FIG, and DiDeMo-FIG demonstrate the effectiveness of EVOVIDEORAG in unifying video retrieval, temporal localization, and question answering through task-adaptive evidence acquisition within a unified agentic framework.

1 INTRODUCTION

With the emergence of multimodal foundation models, Retrieval-Augmented Generation (RAG) has expanded from text to documents, images, audio, and video (Chen et al., 2022; Abootorabi et al., 2025). Video introduces unique challenges: its content is highly redundant, while query-relevant information is often sparse across temporal scales. Answering a query requires sequential decisions over video selection, temporal localization, frame inspection, and termination, while repeated observation incurs substantial visual-token and computational costs (Ye et al., 2025; Buch et al., 2025). To overcome fixed pipelines, Agentic RAG incorporates planning, tool use, and reflection. Self-RAG learns retrieval decisions and generation evaluation, while Adaptive-RAG selects strategies based on question complexity (Asai et al., 2024; Jeong et al., 2024). In video domains, VideoRAG retrieves relevant videos from a corpus, while VideoAgent, DrVideo, LVAgent, and TimeSearch-R

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

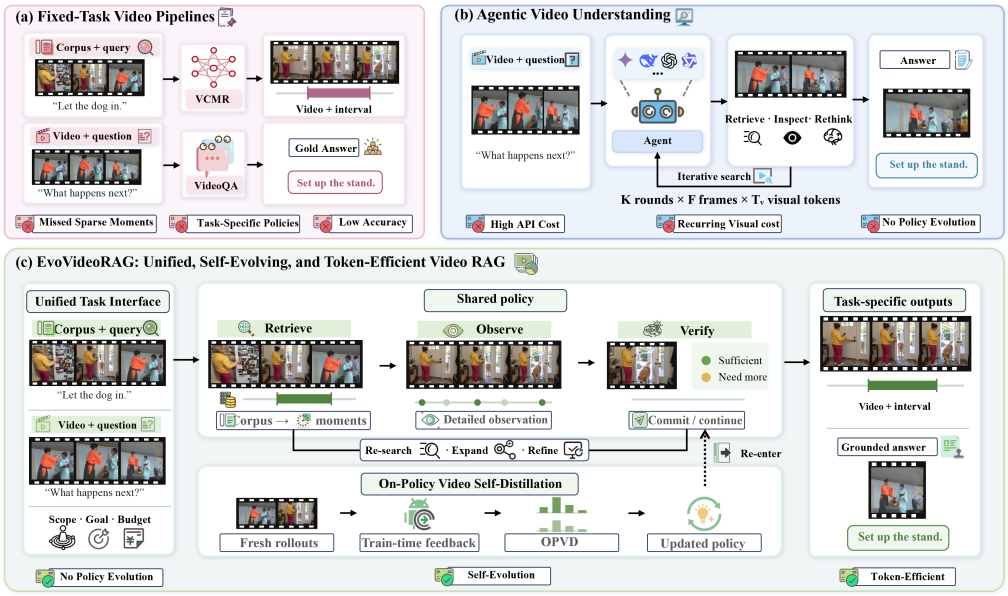


Figure 1: Motivation and overall workflow of EvoVideoRAG.

improve search efficiency through dynamic observation, augmentation, collaboration, and temporal reasoning (Jeong et al., 2025; Wang et al., 2024; Ma et al., 2025; Chen et al., 2025; Pan et al., 2026). However, existing methods mainly target individual video tasks and lack a unified policy for corpus-level retrieval, temporal localization, highlight detection, and VideoQA. Moreover, fixed post-deployment policies limit continual improvement from successful and failed trajectories. External skill storage introduces additional overhead, whereas OPSD and SEED show that trajectory-derived supervision can be internalized into model parameters (Zhao et al., 2026; Wu et al., 2026). Yet, such self-evolution has not been systematically explored for unified and cost-controlled video information seeking. As illustrated in Fig. 1, EvoVideoRAG addresses these challenges by integrating task-adaptive retrieval, evidence verification, and on-policy video self-evolution. To this end, we propose EvoVideoRAG, a unified, task-adaptive, self-evolving, and cost-aware RAG framework for video information seeking. EvoVideoRAG uses a task contract to specify the search scope, content granularity, termination conditions, and output format, from which it constructs a query plan. A Retrieval Navigator performs corpus discovery and temporal localization over a shared video-and-moment index, while a metacognitive policy evaluates the current state through an "observe–rethink–act" loop and autonomously decides whether to continue retrieval, expand the temporal range, increase observation fidelity, or submit the result. To enable continual improvement from its own interactions, we further introduce an On-Policy Video Self-Distillation (OPVD) mechanism for policy self-evolution. The current policy generates new on-policy trajectories; training-time verification and ephemeral procedural guidance for the current round jointly produce fine-grained soft policy signals, which are then distilled into the shared policy through on-policy self-distillation. Our contributions are threefold:

- **Unified agentic framework for heterogeneous Video Understanding.** We provide a unified agentic framework for VCMR and VideoQA, allowing a shared agent policy to adapt to search scopes, content granularities, termination conditions, and output formats.
- **On-Policy Video Self-Distillation for Policy Self-Evolution.** We introduce OPVD to translate evidence gaps and quality–cost feedback into token-level supervision over tool actions and their arguments. Forward-KL distillation integrates targeted correction with preservation of successful, cost-efficient behavior, driving policy self-evolution.
- **Cost-aware Dynamic Visual Budget Allocation.** The system dynamically coordinates retrieval rounds, frame sampling, and observation fidelity according to task demands and evidence sufficiency, directing visual computation toward informative observations to balance task performance and visual-token cost.

2 Related Work

2.1 Multimodal RAG and Video Corpus Retrieval

Traditional textual RAG studies retrieval–generation coordination: selecting useful information from corpora and incorporating it into generation. REALM integrates a learned retriever into pre-training, canonical RAG jointly models retrieval and conditional generation, and Fusion-in-Decoder separately encodes candidate passages before decoder fusion (Guu et al., 2020; Lewis et al., 2020; Izacard & Grave, 2021), establishing the retrieve–fuse–generate pipeline. As retrieval expands beyond text, multimodal RAG investigates how heterogeneous modalities can be represented and matched. MuRAG aligns and retrieves images and text in a shared space before multimodal answering (Chen et al., 2022); VideoRAG uses a video-language model for corpus-level retrieval and answer generation (Jeong et al., 2025). However, existing video RAG systems predetermine retrieval scope, granularity, and objectives, with search and reasoning handled by fixed modules. EvoVideoRAG instead exposes existing video and moment indices as tools, enabling a shared policy to organize corpus search, within-video inspection, and result submission by task conditions.

2.2 Agentic Video Retrieval and Temporal Search

When complex problems cannot be resolved by a single retrieval step, agentic retrieval uses intermediate reasoning to control search, allowing retrieval and reasoning to alternate. ReAct alternates reasoning traces and tool actions, IRCoT generates retrieval requests from chain-of-thought steps, and FLARE determines when additional retrieval is needed using future-content prediction and generation confidence (Yao et al., 2022; Trivedi et al., 2023; Jiang et al., 2023). These methods turn retrieval from a one-shot operation into a reasoning-driven process. Because useful information in long videos is often temporally sparse, video agents decide which segments to inspect instead of processing the full video. VideoAgent, DrVideo, LVAgent, and TimeSearch-R explore tool-based video search through LLM control, document abstraction, multi-agent coordination, and reinforcement learning, respectively (Wang et al., 2024; Ma et al., 2025; Chen et al., 2025; Pan et al., 2026). However, existing methods mainly optimize active search within a single video and lack unified support for open-corpus video selection and fine-grained temporal localization. EvoVideoRAG introduces a task contract and shared state–action interface, enabling one policy to switch among corpus search, video inspection, and different output objectives.

2.3 Experience Internalization and Policy Self-Evolution

Self-evolving agents preserve and reuse experience, failure causes, and corrective strategies from continual interaction. Reflexion converts task feedback into verbal reflections stored in episodic memory, while Voyager packages verified behaviors as executable programs in an expanding skill library (Shinn et al., 2023; Wang et al., 2023). Recent methods compress interaction experience into reusable skills and integrate it with policy optimization. SkillRL extracts a hierarchical Skill Bank from successful and failed trajectories and co-evolves it with the agent policy (Xia et al., 2026). Although external skills preserve interpretable procedural knowledge, their storage and retrieval costs grow with accumulated experience. A complementary direction absorbs experience into policy parameters: OPSD distills policy distributions from visited states back into the model, enabling improvement from self-generated trajectories without external memory (Zhao et al., 2026). However, existing studies do not jointly address corpus-level navigation, temporal localization, and costly visual observation across heterogeneous video tasks. EvoVideoRAG unifies task-conditioned control, policy self-evolution, and dynamic visual budgeting to address these challenges.

3 Method

3.1 Motivation

Video understanding requires finding relevant content and verifying whether it supports reliable localization and reasoning. Retrieved segments may omit event boundaries, temporal relations, or fine-grained details, requiring further search or observation. Queries also vary in search scope and observation requirements, making fixed retrieval and dense sampling susceptible to redundant

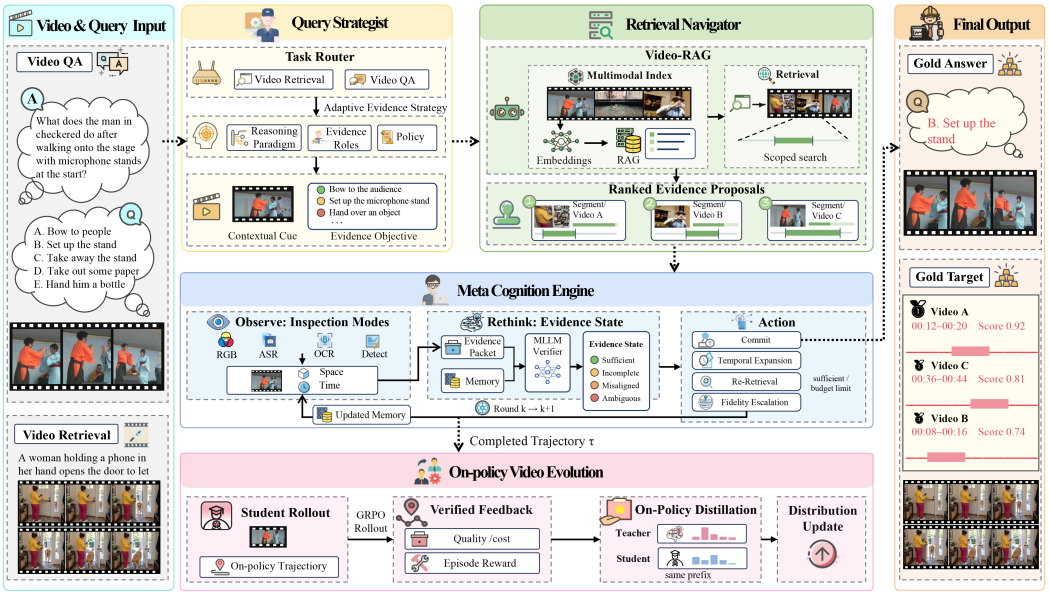


Figure 2: Overview of EVOVIDEORAG. The framework unifies task-adaptive retrieval, evidence verification, and on-policy video self-evolution for efficient video understanding.

computation or missing context. These challenges motivate a unified framework that aligns visual computation with evidence needs and improves decisions through interaction experience. Multi-scale navigation identifies and refines search regions; collaborative verification turns evidence gaps into targeted observations and corrective actions; and on-policy video self-distillation internalizes interaction feedback into the policy. Together, they connect "where to look," "what remains to be understood," and "how to improve next time," linking adaptive inference with policy self-evolution during training.

3.2 SYSTEM OVERVIEW

As shown in Fig. 2, EVOVIDEORAG integrates a Query Strategist, Retrieval Navigator, Meta Cognition Engine, and On-Policy Video Evolution module into a unified agentic RAG framework. Given a query q and video collection Ω, it returns ranked video–moment pairs or answers grounded in acquired evidence. Within each task, functional roles share the task-specific policy weights.

The Query Strategist identifies reference events and temporal relations, while the Retrieval Navigator retrieves candidate videos and refines temporal regions. The Meta Cognition Engine maintains Temporal Evidence Memory and verifies evidence relevance, completeness, and sufficiency to guide further retrieval, observation refinement, or submission.

During training, On-Policy Video Self-Distillation (OPVD) converts evidence gaps and quality–cost feedback into self-teacher supervision. Token-level forward-KL distillation refines tool actions and their arguments while preserving successful, cost-efficient behavior. The updated policy collects fresh trajectories, closing the self-evolution loop. At inference, the learned policy controls evidence acquisition and the frozen QA answerer produces answers, without guidance.

3.3 MULTISCALE NAVIGATION AND COLLABORATIVE VERIFICATION

Given a query q, EVOVIDEORAG alternates between navigation, observation, and verification to localize relevant moments or answer questions. Temporal windows are indexed using a retrieval encoder. For corpus-level retrieval, Window-max ranks each video by its highest query–window similarity, as evaluated in Section 4.2. The Query Strategist identifies reference events and temporal relations, guiding the Retrieval Navigator from coarse event regions to local segments. Candidate

selection balances relevance and temporal coverage while suppressing overlapping windows:

$$\mathcal{C}_0 = \{(v_i, [l_i, r_i], r_i^{\text{rel}})\}_{i=1}^{K},$$

where v_i identifies a video, $[l_i, r_i]$ is a candidate interval, r_i^{rel} is its relevance score, and K is the candidate count.

The Meta Cognition Engine coordinates observation, verification, and decision-making through temporal memory $\mathcal{M}_t$. Each memory unit records a temporal interval, event description, and source, drawing on visual observations and available text within the retrieved window. Verification assesses relevance, temporal completeness, and information sufficiency to guide the policy:

$$\begin{aligned} z_t &= \mathcal{V}(q, \mathcal{C}_t, \mathcal{M}_t), \qquad a_t \sim \pi_\theta(\cdot \mid s_t), \\ s_t &= (q, \mathcal{C}_t, \mathcal{M}_t, z_t). \end{aligned} \tag{1}$$

Here, $\mathcal{V}$ produces assessment z_t, and π_θ controls navigation and observation. Mismatches trigger renewed retrieval; missing context prompts directional temporal expansion; ambiguous details call for denser sampling or higher resolution; and sufficient evidence leads to submission. Candidate deduplication and memory reuse focus further observation on unresolved evidence gaps.

At termination, localization returns verified video–moment pairs, while QA uses a frozen answerer:

$$\hat{y} = F_\phi(q, E_T), \tag{2}$$

where E_T contains the final temporal memory and key observations, and ϕ remains fixed during policy training. States, actions, observations, and computational costs are recorded in trajectory τ for subsequent policy evolution.

3.4 Feedback-Driven Policy Self-Evolution

On-Policy Video Self-Distillation (OPVD) refines retrieval, observation, and stopping through evidence-acquisition feedback. Each round takes the current policy π_k, training queries, the video index, and the fixed QA answerer F_ϕ, and returns π_{k+1}, linking evidence gaps, cost-aware action evaluation, and tool-call distillation.The full training procedure is summarized in Appendix A.4.

Evidence-conditioned action feedback. The current policy collects fresh trajectories $\mathcal{D}_k$ with the evidence states in Eq. equation 1. At mismatched, incomplete, ambiguous, or proposed stopping states, we compare the recorded action with legal alternatives from retrieval candidates, temporal neighbors, finer observation settings, and submission. Each action is evaluated using one sampled continuation under the frozen current policy:

$$U_s(a) = R_d(\tau_{s,a}) - \lambda \widetilde{C}_{\text{vis}}(\tau_{s,a}), \tag{3}$$

where d denotes the task. Using training annotations, R_d is exact option-match accuracy for QA and $\mathbb{1}[\hat{v} = v^\star]$ $\text{IoU}(\hat{I}, I^\star)$ for localization. Visual cost sums input visual tokens across all continuation calls, including repeated historical frames and the fixed QA answerer. It is normalized by the mean full-trajectory visual cost for the same task in $\mathcal{D}_k$, with the denominator bounded below by one. Thus, λ provides soft control over task quality and observation cost.

From evidence gaps to temporary guidance. Let $a_s^\star$ be the evaluated action with the highest utility. If it improves upon the recorded action, a fixed template forms guidance h_s from the verification assessment, preferred action and arguments, and reward–cost comparison. This guidance specifies the evidence gap and corrective operation without supplying the ground-truth answer, video identity, or temporal boundaries. A frozen policy snapshot produces the self-teacher distribution:

$$\mu_k(\cdot \mid s) = \bar{\pi}_k(\cdot \mid s, h_s). \tag{4}$$

The student receives only s, internalizing feedback-informed corrections into decisions based on observed evidence.

Decision distillation and behavior preservation. The correction set $\mathcal{D}_k^{\text{corr}}$ contains states with a strictly better evaluated action. The preservation set $\mathcal{D}_k^{\text{keep}}$ contains remaining evaluated states from successful trajectories with visual cost at or below the within-task median of successful trajectories. Success requires a correct QA option, or a correct video with temporal IoU ≥ 0.5. The policy update minimizes

$$\mathcal{L}_{\text{OPVD}} = \mathbb{E}_{s\sim\mathcal{D}_k^{\text{corr}}} D_{\text{KL}}(\mu_k(\cdot \mid s) \,\|\, \pi_\theta(\cdot \mid s)) + \beta\, \mathbb{E}_{s\sim\mathcal{D}_k^{\text{keep}}} D_{\text{KL}}(\bar{\pi}_k(\cdot \mid s) \,\|\, \pi_\theta(\cdot \mid s)) \,. \tag{5}$$

The distributions are defined over tool calls. Forward KL is estimated by summing token-level KL at teacher-sampled prefixes over action names and arguments, masking prompt and answer tokens. Distillation therefore targets both operations and execution parameters, including temporal direction and observation fidelity. When both sets are nonempty, we sample uniformly within each and use equal proportions; an empty set contributes zero loss, and sampling continues if both are empty. After each update, guidance is discarded and π_{k+1} collects fresh trajectories, internalizing evidence-specific corrections without an additional teacher or persistent skill library at inference. A representative evidence-verification prompt is provided in Appendix A.8.

4 Experiments

4.1 Experimental Setup

We evaluate EvoVideoRAG on Charades-FIG and DiDeMo-FIG (Chen et al., 2024) for VCMR, and NExT-QA (Xiao et al., 2021) and Video-MME (Fu et al., 2025) for VideoQA. Dataset details are provided in Appendix A.1. Qwen3-VL-Embedding-2B indexes multiscale windows for corpus-wide localization and within-video QA retrieval. Task-specific Qwen3-VL-8B-Instruct policies use LoRA ($r = 16$, $\alpha = 32$), trained with AdamW and BF16 on two RTX 5090 GPUs (32 GB each), with vLLM for trajectory generation and inference. Additional training hyperparameters are provided in Appendix A.3. After SFT, OPVD combines quality–cost feedback with forward-KL distillation and behavior preservation. Inference uses the updated policy and frozen QA answerer, without training-time guidance or a self-teacher. Baselines include closed-source models, VideoMind (Liu et al., 2026), Video-RAG (Luo et al., 2026), TimeSearch-R (Pan et al., 2026), DeepSeek-VL2-Tiny (Wu et al., 2024), InternVL3-8B (Zhu et al., 2025), LLaVA-1.5-7B, and Qwen3-VL-8B (Bai et al., 2025). VCMR@τ measures top-1 success with the correct video and IoU $\geq \tau$, where $\tau \in \{0.3, 0.5, 0.7\}$. E2E mIoU averages IoU across all queries, counting wrong-video predictions and failures as zero. We report visual-token usage and vary input resolution with EvoVideoRAG-Lite to assess the quality–cost trade-off.

4.2 Retrieval Effectiveness

Video retrieval. We evaluate Window-max retrieval with 8-second windows and a 4-second stride, following sliding-window moment matching. Table 1 shows improved recall over whole-video ranking and higher R@10 and R@100 than the published VideoSearch-R1 results on both datasets (Lee et al., 2026). Details of RAG index construction and retrieval analysis are provided in Appendix A.2.

Table 1: Video retrieval recall (R@K, %). VideoSearch-R1 values are retained as published.

Retrieval method	R@1	R@5	R@10	R@100
Charades-FIG				
Whole-video ranking	*12.07*	*25.46*	*32.82*	*55.67*
VideoSearch-R1	**24.6**	42.9	52.2	84.7
Ours	24.46	**45.99**	**56.24**	**87.37**
DiDeMo-FIG				
Whole-video ranking	*41.34*	*62.42*	*68.85*	*82.58*
VideoSearch-R1	**59.0**	**82.0**	87.8	97.5
Ours	57.21	80.74	**89.48**	**97.63**

4.3 MAIN RESULTS

The following tables present the comparison. For direct inference with closed-source APIs and open-source vision–language models, we uniformly sample 16 frames per query on Charades-FIG, DiDeMo-FIG, and NExT-QA, and 32 frames on Video-MME. EVOVIDEORAG acquires evidence through adaptive multi-round observation. On Video-MME, all baselines use subtitles; our agent selects available subtitle entries within each retrieved temporal window.

Video corpus moment retrieval. In the comparison in Table 2, EVOVIDEORAG achieves the highest VCMR@0.3, VCMR@0.7, and E2E mIoU on Charades-FIG, reaching 22.10%, 8.98%, and 14.69%, respectively. Its VCMR@0.5 reaches 16.91%, exceeding Qwen3-VL-8B by 7.90 percentage points. On DiDeMo-FIG, EVOVIDEORAG records the highest VCMR@0.3, VCMR@0.5, and E2E mIoU among the compared methods, at 32.99%, 29.73%, and 19.38%, respectively. Its VCMR@0.5 exceeds the strongest baseline by 13.48 percentage points. Qualitative comparisons and a detailed localization case study are provided in Appendices A.5 and A.7, respectively.

Table 2: Main results on video corpus moment retrieval.

Method	VCMR@0.3 ↑	VCMR@0.5 ↑	VCMR@0.7 ↑	E2E mIoU ↑	Visual Tokens ↓
Charades-FIG					
Closed-Source Models					
GPT-4o	20.22	14.25	5.83	13.47	1157.7
GPT-4.1	19.46	12.20	6.21	13.24	1343.7
Gemini-3-Pro	<u>22.07</u>	**17.23**	<u>8.95</u>	<u>14.68</u>	2356.5
Claude-Sonnet-5	19.87	13.44	5.89	13.09	2444.4
Open-Source Models and Video Agents					
VideoMind	7.02	4.03	1.45	4.46	**457.87**
Video-RAG	7.28	5.11	1.69	4.54	1956.59
TimeSearch-R	12.58	12.50	0.16	8.45	*1008.78*
DeepSeek-VL2-Tiny	6.75	4.62	2.12	4.60	1742.1
InternVL3-8B	7.61	4.35	1.37	5.15	1388.9
LLaVA-1.5-7B	1.56	0.81	0.16	1.17	<u>576.3</u>
Qwen3-VL-8B	14.84	9.01	3.68	9.48	1,418.5
Ours					
EvoVideoRAG	**22.10**	<u>16.91</u>	**8.98**	**14.69**	2,437.8
EvoVideoRAG-Lite	*21.77*	*15.38*	*8.58*	*13.81*	1,502.7
DiDeMo-FIG					
Closed-Source Models					
GPT-4o	18.04	7.45	2.32	13.51	1367.6
GPT-4.1	20.63	8.82	3.39	15.13	1204.4
Gemini-2.5-Pro	27.49	*16.25*	<u>6.23</u>	*18.04*	1569.2
Claude-Sonnet-5	24.27	12.61	4.66	16.88	2490.7
Open-Source Models and Video Agents					
VideoMind	12.41	4.76	1.89	9.36	**448.08**
Video-RAG	19.14	14.08	1.89	11.56	1958.40
TimeSearch-R	<u>31.92</u>	6.13	0.37	15.53	1200.36
DeepSeek-VL2-Tiny	13.21	7.10	1.47	8.62	1906.5
InternVL3-8B	9.67	3.59	0.77	6.55	2221.4
LLaVA-1.5-7B	2.27	0.62	0.15	2.75	<u>576.8</u>
Qwen3-VL-8B	17.72	8.72	3.76	12.24	1,471.0
Ours					
EvoVideoRAG	**32.99**	**29.73**	*5.86*	**19.38**	1382.5
EvoVideoRAG-Lite	*30.73*	<u>28.06</u>	**7.77**	<u>18.67</u>	*1,079.4*

Video question answering. In the comparison in Tables 3 and 4, EVOVIDEORAG achieves the highest overall accuracies among the compared methods, at 78.44% on NExT-QA and 69.26% on

Video-MME. On the short-, medium-, and long-video subsets of Video-MME, its accuracies are 75.56%, 70.56%, and 61.67%, respectively. On the medium-video subset, it exceeds Gemini-3-Pro by 1.67 percentage points. A detailed QA case study is presented in Appendix A.6.

Table 3: Main results on NExT-QA test split. C/T/D: causal, temporal, and descriptive questions.

Method	Acc@C ↑	Acc@T ↑	Acc@D ↑	Acc@All ↑	Visual Tokens ↓
Closed-Source Models					
GPT-4o	66.79	62.66	66.12	65.40	1258.8
GPT-4.1	**80.52**	**75.84**	75.59	78.26	*1104.6*
Gemini-3-Pro	77.61	74.82	*84.06*	77.80	1767.2
Claude-Sonnet-5	62.79	59.28	74.38	63.60	2412.1
Open-Source Models and Video Agents					
VideoMind	78.03	74.82	84.13	*78.04*	1651.5
Video-RAG	73.81	65.98	78.58	72.16	1982.3
TimeSearch-R	79.30	66.84	80.71	75.67	1851.98
DeepSeek-VL2-Tiny	55.89	55.48	61.64	56.70	1935.9
InternVL3-8B	55.40	46.78	52.17	52.20	2330.6
LLaVA-1.5-7B	57.55	49.49	61.35	55.67	**576.8**
Qwen3-VL-8B	73.26	66.92	78.29	72.12	3964.5
Ours					
EvoVideoRAG	*79.08*	*74.18*	**84.48**	**78.44**	1138.6
EvoVideoRAG-Lite	78.23	74.03	83.77	77.84	802.9

Table 4: Main results on Video-MME.

Method	Short ↑	Medium ↑	Long ↑	Acc@All ↑	Visual Tokens ↓
Closed-Source Models					
GPT-4o	61.67	44.44	44.44	50.19	*1021.3*
GPT-4.1	71.11	64.44	58.89	64.81	884.7
Gemini-3-Pro	76.11	68.89	62.22	69.07	2254.9
Claude-Sonnet-5	66.11	56.67	57.78	60.19	2912.4
Open-Source Models and Video Agents					
VideoMind	67.22	52.78	48.33	56.11	2518.4
Video-RAG	67.78	66.11	**63.33**	65.74	3360.7
TimeSearch-R	55.00	48.33	32.78	45.37	2089.0
DeepSeek-VL2-Tiny	53.33	40.56	33.33	42.41	2012.4
InternVL3-8B	46.67	40.56	41.67	42.96	1268.3
LLaVA-1.5-7B	46.67	33.89	33.89	38.15	**573.9**
Qwen3-VL-8B	70.56	60.56	53.33	61.48	1952.1
Ours					
EvoVideoRAG	*75.56*	**70.56**	*61.67*	**69.26**	2477.5
EvoVideoRAG-Lite	**76.67**	*68.33*	60.00	*68.33*	2013.2

Token efficiency. On benchmarks, EvoVideoRAG is competitive with closed-source APIs at visual-token usage within their range, while outperforming several lower-token open-source video agents in E2E mIoU. Figure 3(a) shows that EvoVideoRAG-Lite further reduces visual-token usage while retaining most of the full model's performance.

4.4 Ablation Study

Tables 5 and 6 compare the SFT-initialized Agentic Base, GRPO, and EvoVideoRAG. For each task, GRPO and OPVD share SFT initialization, training data, rewards, tool interfaces, and evaluation

protocols. Over GRPO, EvoVideoRAG improves VCMR@0.5 by 2.61 and 4.89 percentage points on Charades-FIG and DiDeMo-FIG, respectively. Overall QA accuracy improves by 3.26 points on NExT-QA and 9.26 points on Video-MME, including an 11.11-point gain on long videos. Gains are computed before rounding.

Table 5: Ablation on video corpus moment retrieval. All metrics are percentages.

Variant	VCMR@0.3 ↑	VCMR@0.5 ↑	VCMR@0.7 ↑	E2E mIoU ↑
Charades-FIG				
Agentic Base	*17.88*	*13.36*	*5.19*	*11.53*
GRPO	19.76	14.30	6.40	12.90
EvoVideoRAG	**22.10**	**16.91**	**8.98**	**14.69**
DiDeMo-FIG				
Agentic Base	*26.74*	*19.56*	*2.99*	*16.09*
GRPO	27.46	24.84	5.61	17.46
EvoVideoRAG	**32.99**	**29.73**	**5.86**	**19.38**

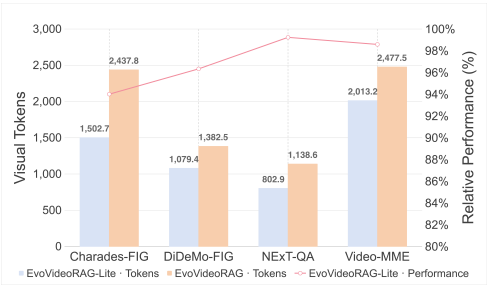
Table 6: Ablation on video question answering. All metrics are accuracy percentages. C, T, D denote causal, temporal, and descriptive questions; S, M, L denote short, medium, and long videos.

(a) NExT-QA

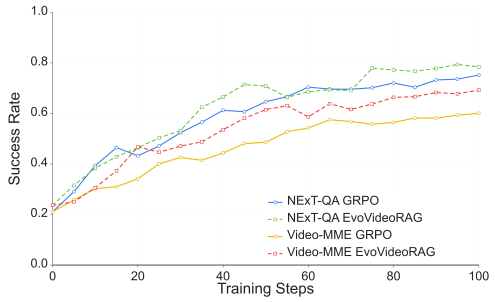
Variant	C ↑	T ↑	D ↑	All ↑
Agentic Base	*73.97*	*66.54*	*78.58*	*72.42*
GRPO	75.41	71.25	81.85	75.18
EvoVideoRAG	**79.08**	**74.18**	**84.48**	**78.44**

(b) Video-MME

Variant	S ↑	M ↑	L ↑	All ↑
Agentic Base	*71.67*	*57.22*	*50.00*	*59.63*
GRPO	72.78	56.67	50.56	60.00
EvoVideoRAG	**75.56**	**70.56**	**61.67**	**69.26**



(a) Token efficiency



(b) Policy evolution

Figure 3: Token efficiency and policy evolution.

Policy evolution. Figure 3(b) reports validation success rates during training. EvoVideoRAG shows an upward trend and higher late-stage success rates than GRPO on both benchmarks.

5 Conclusion

We present EvoVideoRAG, a unified framework for adaptive and efficient video information seeking. By modeling video understanding as sequential evidence acquisition, it enables a shared policy for retrieval, verification, and observation control. Our on-policy video distillation further improves the policy from interaction trajectories without persistent external memory. Experiments across retrieval and QA benchmarks demonstrate effective and cost-aware video understanding.

AI USE STATEMENT

Generative AI tools were used for language editing, improving clarity and readability, and providing suggestions for manuscript organization. The authors are responsible for all scientific content, methodological decisions, experimental procedures, data analyses, and the final manuscript.

ETHICS STATEMENT

This work studies video understanding using publicly available benchmarks. To the best of our knowledge, this work raises no ethical concerns requiring special consideration.

REPRODUCIBILITY STATEMENT

The main text and appendices provide experimental details, including dataset composition (Figure 4), RAG index construction (Appendix A.2), and training hyperparameters (Appendix A.3). We will publicly release our code and configurations upon acceptance.

REFERENCES

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16776–16809, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

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

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.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20237–20246. IEEE, 2025.

Houlun Chen, Xin Wang, Hong Chen, Zeyang Zhang, Wei Feng, Bin Huang, Jia Jia, and Wenwu Zhu. Verified: A video corpus moment retrieval benchmark for fine-grained video understanding. *Advances in Neural Information Processing Systems*, 37:40393–40406, 2024.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, 2022.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pp. 874–880, 2021.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 7036–7050, 2024.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298, 2025.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7969–7992, 2023.

Seohyun Lee, Seoung Choi, Dohwan Ko, Jongha Kim, and Hyunwoo J Kim. Videosearch-r1: Iterative video retrieval and reasoning via soft query refinement. In *European Conference on Computer Vision*, pp. 648–664. Springer, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

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, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946. IEEE, 2025.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8376–8405, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 10014–10037, 2023.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, et al. Seed: Self-evolving on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.14777*, 2026.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

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.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

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.

A Appendix

A.1 Dataset Overview

The datasets used in this work cover two representative categories of video-language understanding tasks: video corpus moment retrieval, including Charades-FIG and DiDeMo-FIG, and video question answering, including NExT-QA and Video-MME. These benchmarks are selected to evaluate whether a single evidence acquisition policy can generalize across different search scopes, temporal granularities, and reasoning objectives.

Specifically, video corpus moment retrieval requires locating relevant moments from a large video collection according to textual descriptions. Such tasks emphasize two sequential decisions: identifying the correct video among many candidates and refining the temporal interval containing the described event. In contrast, video question answering focuses on extracting sufficient evidence from a given video to answer questions involving temporal, causal, or descriptive reasoning. These tasks therefore stress different aspects of evidence acquisition: corpus-level retrieval, fine-grained temporal localization, and within-video reasoning.

The selected benchmarks provide complementary evaluation scenarios. Charades-FIG and DiDeMo-FIG contain natural-language descriptions paired with target moments, enabling evaluation of video selection and boundary recovery. NExT-QA evaluates whether an agent can acquire evidence required for reasoning-oriented questions, while Video-MME further introduces longer videos and diverse temporal durations to examine evidence acquisition under extended contexts. Together, these datasets cover short- and long-range temporal reasoning, open-corpus search, and fixed-video question answering.

Figure 4 summarizes their annotation formats, target interval durations, and question categories, highlighting the diversity of temporal structures and evidence requirements covered by the evaluation suite. The diversity of these benchmarks allows us to evaluate whether the proposed framework learns a general evidence acquisition policy rather than optimizing for a single task-specific retrieval procedure. The figure summarizes dataset-level characteristics across released splits and is intended to illustrate the diversity of annotation formats, temporal structures, and reasoning requirements.

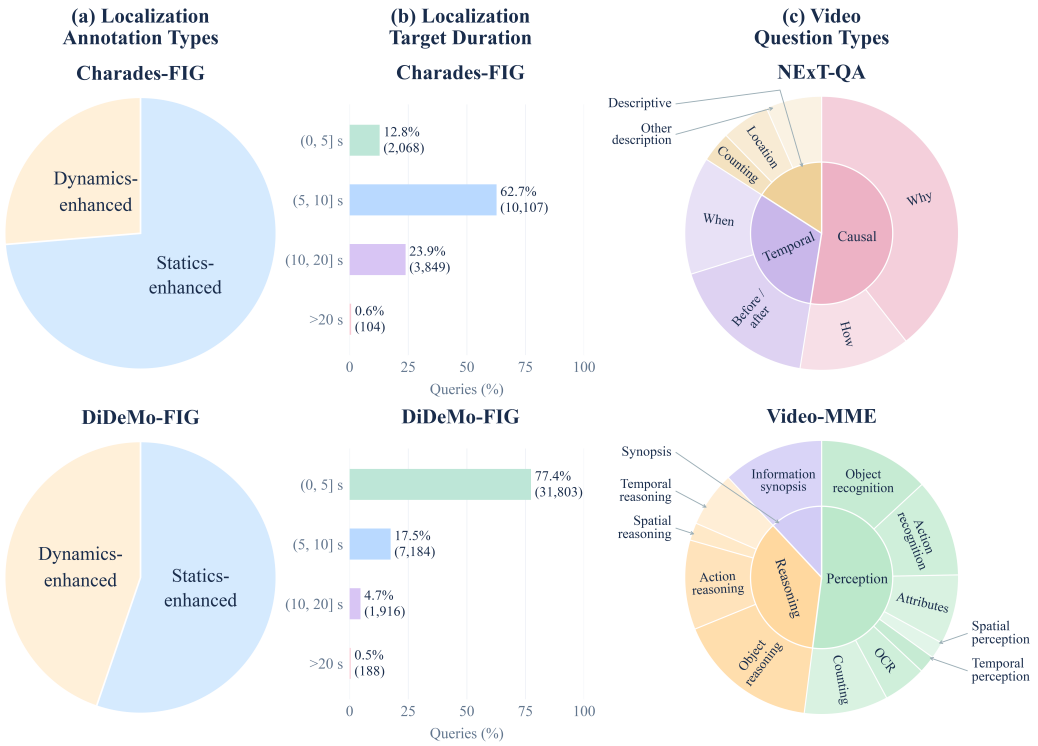


Figure 4: Dataset composition and task characteristics: (a) localization annotation formats; (b) target temporal interval distributions; (c) question categories and reasoning types.

The constructed indices support both corpus-level retrieval and within-video evidence inspection. For localization tasks, the agent first searches over video-level candidates and then refines relevant temporal regions using indexed windows. For question answering tasks, retrieval is restricted to the provided video while using the same temporal representation. This shared indexing strategy allows different video understanding tasks to operate through a unified evidence acquisition interface.

A.2 RAG Index Construction and Retrieval Analysis

Index coverage across training and evaluation. We build offline RAG indices using Qwen3-VL-Embedding-2B with 2,048-dimensional representations, providing reusable evidence access for trajectory collection and evaluation. Table 7 summarizes the revised partitions and temporal-unit counts. Charades-FIG reserves approximately 10% of its official training videos for validation while retaining the official test set. DiDeMo-FIG and NExT-QA follow their official splits. Video-MME uses a custom, duration-stratified 60/20/20 train/validation/test split. The temporal units in Table 7 correspond to indexed retrieval candidates rather than independent training samples. Each video is decomposed into multiscale temporal windows, which serve as searchable evidence units during navigation. Therefore, the number of temporal units reflects the coverage and granularity of the retrieval space rather than the number of annotated queries. Maintaining video-disjoint partitions ensures that retrieval candidates observed during training do not overlap with evaluation videos, preventing information leakage during both trajectory collection and final assessment. All revised partitions are video-disjoint.

Table 7: Dataset partitions and offline retrieval index statistics.

Dataset	Split	Videos	Queries	Temporal units
Charades-FIG	train	4,804	11,122	34,690
	val	534	1,286	3,738
	test	1,334	3,720	9,154
	Total	**6,672**	**16,128**	**47,582**
DiDeMo-FIG	train	8,368	32,903	98,212
	val	1,064	4,175	12,024
	test	1,002	4,013	12,035
	Total	**10,434**	**41,091**	**122,271**
NExT-QA	train	3,870	34,132	41,618
	val	570	4,996	5,743
	test	1,000	8,564	9,396
	Total	**5,440**	**47,692**	**56,757**
Video-MME	train	540	1,620	68,613
	val	180	540	23,155
	test	180	540	23,576
	Total	**900**	**2,700**	**115,344**

A.3 Training Hyperparameters

We provide the main training hyperparameters for both supervised fine-tuning and agent policy self-evolution in Table 8. The two training stages share the same LoRA configuration and optimizer type, while using different learning rates according to their objectives. Supervised fine-tuning focuses on establishing basic tool-use and evidence-acquisition capabilities, whereas policy self-evolution further optimizes retrieval, observation, and decision behaviors through on-policy distillation. Effective batch sizes account for gradient accumulation and data parallelism across GPUs. During supervised fine-tuning, the policy learns the basic interaction protocol, including retrieval, observation, and submission behaviors. During self-evolution, the policy is further optimized using on-policy trajectories and distilled feedback, where a smaller learning rate is adopted to preserve previously acquired capabilities while incorporating new interaction experience.

Table 8: Main training hyperparameters of EvoVideoRAG. Effective batch sizes include gradient accumulation and data parallelism.

Hyperparameter	Setting
Shared configuration	
LoRA rank / scaling / dropout	16 / 32 / 0.05
Learning-rate schedule / weight decay	Cosine / 0.01
Supervised fine-tuning	
Learning rate	1×10^{-4}
Effective batch size	8
Agent policy self-evolution	
Learning rate	1×10^{-6}
Effective batch size	8
Training steps	100

All experiments use the same backbone architecture and LoRA adaptation configuration across training stages. The difference between the two stages lies in the optimization objective rather than the model structure: supervised fine-tuning establishes initial evidence acquisition behaviors, whereas policy self-evolution improves decision quality through additional interaction trajectories. The reported effective batch size includes both gradient accumulation and multi-GPU data parallelism.

A.4 ALGORITHMIC OVERVIEW

We summarize the complete training and inference procedure of EVOVIDEORAG in Algorithm 1. The procedure consists of two interconnected stages: evidence-driven interaction and policy self-evolution. Starting from an SFT-initialized policy, the agent interacts with indexed video collections by constructing retrieval candidates, maintaining temporal evidence memory, and selecting actions for further retrieval, temporal expansion, observation refinement, or final commitment. These interactions produce trajectories containing states, decisions, and visual costs.

At each evolution round, the current policy generates on-policy trajectories, after which visited states are evaluated according to task utility and visual cost. Corrective behaviors and successful behaviors are separated into two sets and used to construct temporary guidance for self-teacher action distributions. The policy is then updated through on-policy video distillation while preserving existing capabilities. The temporary guidance is removed after each update, enabling continual improvement without requiring a persistent skill library during inference.

Algorithm 1 OPVD training procedure of EVOVIDEORAG.

Input: SFT policy π_0, labeled training queries $\mathcal{Q}$, video index Ω, fixed QA answerer F_ϕ, rounds L
Output: Evolved policy π_L

for $k = 0, \ldots, L-1$ **do**
  Freeze policy snapshot $\bar{\pi}_k \leftarrow \pi_k$
  Collect fresh trajectories $\mathcal{D}_k$ through retrieval, observation, and verification; record task outputs and visual costs
  Select states with evidence gaps or proposed termination, and construct their legal candidate actions
  Sample one continuation per candidate under $\bar{\pi}_k$; evaluate $U_s(a)$ using Eq. equation 3
  Construct $\mathcal{D}_k^{\text{corr}}$ and $\mathcal{D}_k^{\text{keep}}$ using the utility-improvement and success-cost criteria
  for all $s \in \mathcal{D}_k^{\text{corr}}$ **do**
    Construct temporary guidance h_s from evidence gaps and action-utility comparisons
    Obtain self-teacher distribution $\mu_k(\cdot \mid s) = \bar{\pi}_k(\cdot \mid s, h_s)$
  end for
  Initialize $\pi_{k+1} \leftarrow \pi_k$
  if $\mathcal{D}_k^{\text{corr}} \cup \mathcal{D}_k^{\text{keep}} \neq \emptyset$ **then**
    Update policy adapters using Eq. equation 5, keeping F_ϕ fixed and h_s hidden from the student
  end if
  Discard temporary guidance and proceed to the next round
end for
return π_L

A.5 QUALITATIVE ANALYSIS.

Figure 5 provides a qualitative comparison of video selection and temporal localization behaviors between EVOVIDEORAG and baseline methods. The examples illustrate typical differences in evidence acquisition: existing approaches retrieve semantically related but incorrect videos or identify incomplete temporal regions when the target event requires reasoning over multiple actions and temporal contexts.

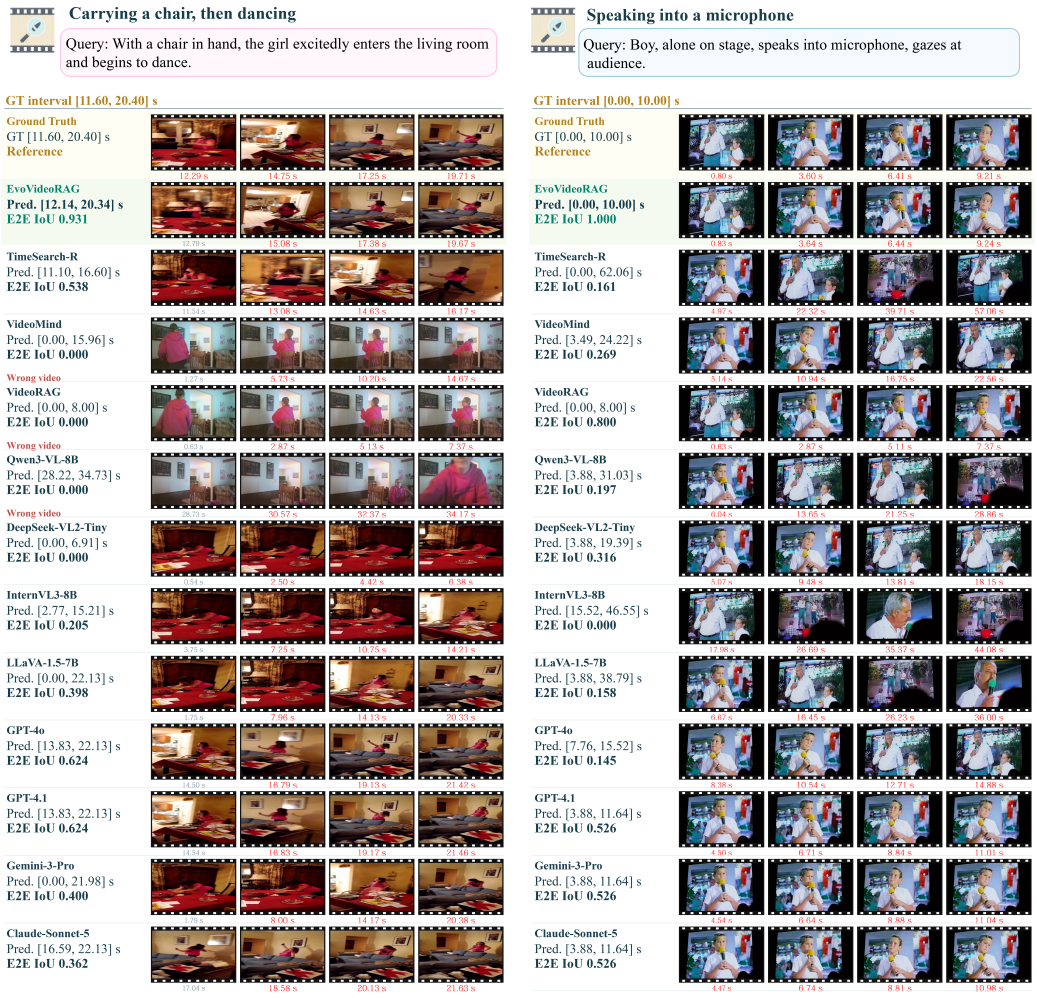


Figure 5: Qualitative comparison on video corpus moment retrieval.

Compared with these methods, EvoVideoRAG performs evidence acquisition through an iterative process of retrieval, observation, and verification. Instead of directly committing to the highest-ranked retrieval result, the agent evaluates whether the current evidence satisfies the temporal and semantic requirements of the query. When retrieved evidence is incomplete, it can switch candidates, expand temporal regions, or increase observation fidelity to resolve remaining uncertainty. This behavior enables more accurate recovery of event boundaries and reduces errors caused by visually similar but semantically incorrect content.

For sequential actions such as "carrying a chair, then dancing," EvoVideoRAG captures both parts of the event instead of focusing on only one action or a visually similar distractor. This indicates that the agent can preserve temporal relationships between multiple actions rather than matching isolated visual concepts. For fine-grained actions such as "speaking into a microphone," it further demonstrates the ability to identify the relevant moment without unnecessary temporal expansion, showing that the adaptive observation strategy can also avoid excessive search when sufficient evidence has already been acquired.

A.6 Case Study: Evidence-Guided Temporal Expansion

We present a NExT-QA example in which the agent identifies missing temporal evidence and inspects a later neighboring window. The accumulated evidence then supports answering by the frozen

QA answerer. **Input. Video:** A 55.5-second skiing video.
Question: "What did both of them do after completing skiing?"
Options: A. Jump and pose; B. Bend down; C. Lift and hold by the side; D. Raised their hands; E. Make turn.

(1) Initial retrieval: `SEARCH`. **Input:** The question and the index of the specified video.
Action: Retrieve five candidate intervals using the question.
Output: The top-ranked candidate is $[28, 36]$ seconds, which serves as the initial observation window.

(2) Initial observation: `INSPECT`. **Input:** The question and the retrieved interval $[28, 36]$ seconds.
Action: Inspect eight frames from this interval.
Output: The observations establish the skiing action, but the subsequent event remains unresolved. The agent records `PARTIAL_AFTER` and identifies `event_after_anchor` as the missing evidence slot. Figure 6 shows representative observations.



28.15 s 30.35 s 35.85 s

Figure 6: Representative frames from the initial observation window, $[28, 36]$ seconds.

(3) Directional temporal expansion: `EXPAND`. **Input:** The initial interval and the assessment that the event after skiing remains unresolved.
Action: Select the next temporal neighbor using `direction=after` and `hops=1`.
Output: A later overlapping window, $[32, 40]$ seconds, extending the region available for subsequent observation. This step returns candidate information without acquiring new frames.

(4) Additional observation: `INSPECT`. **Input:** The question, accumulated evidence, and the neighboring interval $[32, 40]$ seconds.
Action: Inspect eight frames from this window and update the temporal evidence memory.
Output: The observations show the two skiers raising their hands after skiing, including the frames at 37.65 and 39.85 seconds. The agent marks the evidence as `SUFFICIENT` and proceeds to answering. Figure 7 shows the additional observations.



35.45 s 37.65 s 39.85 s

Figure 7: Additional observations from $[32, 40]$ seconds reveal the action after skiing.

(5) Evidence-grounded answer: `ANSWER`. **Input:** The question, answer options, and accumulated evidence E_T, comprising temporal memory and key observations.

Action: Pass the acquired evidence to the frozen QA answerer F_ϕ to generate the final answer.
Output: D, "Raised their hands," supported by observations from $[32, 40]$ seconds. The answer matches the dataset annotation.

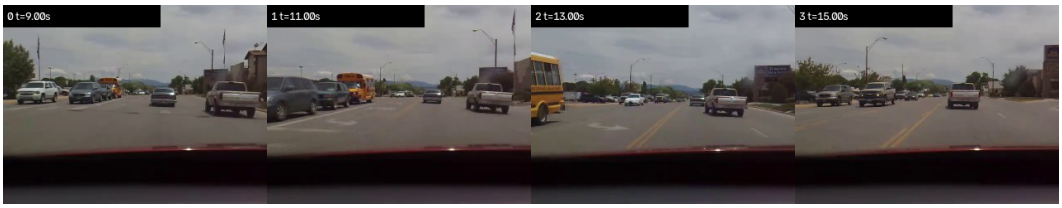
A.7 Case Study: Corpus-Level Retrieval and Localization

We present a DiDeMo-FIG example in which EvoVideoRAG corrects an initial video selection and verifies the target event through local inspection. While the preceding QA example searches within a specified video, this case requires selecting a video from the corpus and identifying its relevant temporal interval.

Input. Task: Retrieve and localize the described event in the video corpus.
Query: *A yellow school bus travels from right to left with a stationary black cat observing from a red couch.*

(1) Initial observation and candidate diagnosis. Input: The query and the initial candidate, Video A, at $[8, 16]$ seconds.
Action: Inspect four frames from the candidate interval.
Output: The frames contain road traffic and a yellow school bus, but do not match the cat-and-couch context. The agent selects `SEARCH_NEXT` to switch candidates.

(2) Candidate correction and observation. Input: The retrieved candidate list and the preceding evidence assessment.
Action: Switch to Video B using `SEARCH_NEXT`, followed by inspection of four frames from its candidate interval, $[4, 12]$ seconds.
Output: The observations show a dark cat on a red couch beside a window, with a yellow school bus outside. The agent selects `INSPECT_DENSE` to examine the temporal progression more closely. Figure 8 compares the observations before and after candidate correction.



(a) Initial candidate: Video A, $[8, 16]$ s.



(b) Switched candidate: Video B, $[4, 12]$ s.

Figure 8: Recorded coarse observations before and after candidate correction.

(3) Local verification: `INSPECT_DENSE`. Input: The query, accumulated evidence, and Video B's interval, $[4, 12]$ seconds.
Action: Inspect eight frames within the same interval.
Output: The denser sequence shows the bus passing from right to left outside the window while the cat remains on the red couch. With sufficient evidence for the described event, the agent selects `SUBMIT`. Figure 9 shows the dense observations.



Figure 9: Recorded dense observations used to verify the event in Video B.

(4) Localization output: `SUBMIT`. Input: The selected video, candidate interval, and accumulated temporal evidence.
Action: Submit the video–moment prediction using the run's fixed 5-second temporal-grid alignment.
Output: Video B with interval $[5, 10]$ seconds. For this example, the video identity and interval match the annotation, giving a per-query IoU of 1.000.

A.8 PROMPT

The following template summarizes the shared evidence-verification and action-selection principles of EVOVIDEORAG. Task adapters instantiate the inputs, available tools, action names, and output formats.

Structured Evidence Verification and Action Selection

System role.
You are the evidence-acquisition agent for video understanding. Determine where to search, what requires further observation, and when the available evidence supports submission. Base each decision on the task specification and observed evidence.

Current inputs.

- Query and task specification: `{query}`, `{task_specification}`.
- Retrieved candidates and latest observations: `{candidates}`, `{observations}`.
- Temporal evidence memory and action history: `{memory}`, `{history}`.
- Allowed actions, remaining resources, and response format: `{allowed_actions}`, `{remaining_budget}`, `{output_schema}`.

1. Respect the task and search scope.

- Identify the target entities, actions, scene conditions, and temporal relations required by the query.
- Search across videos only when the task permits corpus-level retrieval. Otherwise, remain within the specified video.
- Use only authorized tools and valid arguments. Preserve the association between each video, interval, and evidence source.

2. Observe the available evidence.

- Read the supplied frames chronologically and use available aligned speech or text when relevant. Distinguish directly observed events from prior hypotheses.
- Combine new observations with temporal evidence memory. Retrieval scores and candidate summaries indicate where to look; they do not independently verify the target event.

3. Verify relevance and completeness.

- Check the complete query rather than a single matching object or action. For temporal queries, verify the requested relation, not merely the reference event.
- Determine whether the evidence is sufficient, temporally incomplete, off target, or visually ambiguous. Identify a concrete missing event, relation, boundary, or visual detail before requesting more observations.

4. Select one permitted next action.

- **Retrieve or switch candidates:** Correct a mismatch with the query. Do not abandon a relevant video merely because its temporal boundaries remain uncertain.
- **Expand or refine the interval:** Follow the missing temporal context in the appropriate direction, using only the operations permitted by the task.
- **Increase observation fidelity:** Inspect a relevant region more densely or at higher spatial resolution when this can resolve an identified visual gap.
- **Commit:** Stop when the evidence is sufficient. Submit verified video–moment pairs for localization, or hand the accumulated evidence to the fixed answerer for QA.

5. Control observation cost.

- Reuse evidence memory and retained keyframes. Avoid repeated acquisition of already sufficient evidence.
- Prefer targeted local observations over broad reinspection. Do not escalate merely because budget remains.
- Respect the available frame, tool-call, and fidelity limits. If further observation is unavailable or unhelpful, follow the task's termination rule using existing evidence.

6. Return the required response.
Choose exactly one allowed action and follow `{output_schema}`. Include a concise evidence assessment and action arguments only when requested by that schema. For a restricted action interface, return only the permitted action code. Do not invent observations, video identifiers, or timestamps, and do not add free-form reasoning outside the required response.