

Routing Before Looking: Query-Adaptive Evidence Acquisition for Long-form Video Understanding

Tianyue Wang^{1,2,3} Xuying Wu⁴ Yuxiang Ma⁵ Ruiming Liang² Jiaxuan Kang³
Yanchao Hao^{3†} Zheng Wei³ Leigang Qu^{6*} Haiyun Guo^{2*} Jinqiao Wang^{1,2}

¹School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences

²Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences

³Tencent ⁴Wuhan University ⁵Southeast University ⁶National University of Singapore

wangtianyue25@mails.ucas.ac.cn, leigangqu@gmail.com, haiyun.guo@nlpr.ia.ac.cn

Abstract

Long-form video understanding remains challenging for video agents due to the mismatch between query demands and evidence acquisition strategies. Although recent planning-before-perception methods outperform query-agnostic pipelines, they often rely on a single dominant strategy, either generation-based strategy or retrieval-based strategy, limiting their ability to handle diverse query demands. We propose **Route2Look**, a lightweight and model-agnostic framework for query-adaptive evidence acquisition in long-form video understanding. Route2Look operates in a *Route-Look-Memorize* loop with three tools: *Global Browse* for holistic context, *Temporal Ground* for explicit temporal cues, and *Semantic Retrieve* for semantic search. The core component is a routing policy that dynamically selects evidence acquisition tools based on the query. To build this policy, Route2Look adopts a two-stage design: first distilling the routing skill from differential contrastive analysis between generation-based and retrieval-based trajectories, and then applying the distilled skill with hard routing rules and continue-or-stop criteria during inference. Experiments on challenging long-video benchmarks show that Route2Look achieves state-of-the-art performance while maintaining comparable frame efficiency across datasets and query types. Oracle routing analysis further reveals the potential of query-adaptive evidence acquisition for future long-form video understanding.

1 Introduction

Long-form video understanding has become an important frontier in multimodal intelligence, with applications in multimodal assistants (Qian

et al., 2025; Weerasinghe et al., 2024), embodied agents (Song et al., 2025; Zitkovich et al., 2023), and autonomous systems (Tian et al., 2024; Jiang et al., 2025; Wei et al., 2025). Recent progress in multimodal large language models (MLLMs) (Wang et al., 2024a; Shu et al., 2025; Lin et al., 2026b; Qiang et al., 2025; Wei et al., 2026) and video agent frameworks (Ma et al., 2025; Yan et al., 2026; Yang et al., 2026a; Zhao et al., 2026; Yang et al., 2026c) has enabled models to process increasingly long and complex videos. However, a key problem remains unsolved: the systematic mismatch between query demands and evidence acquisition strategies.

Based on observations, we identify three distinct query types in real-world scenarios as illustrated in Fig. 1: (1) global understanding queries, such as summarization, main theme identification, or overall video categorization; (2) explicit temporal queries, where the query contains timestamps, time ranges, or temporal expressions; and (3) implicit temporal queries, a common and challenging type where the relevant evidence is not explicitly localized by the query. For the first two types, evidence acquisition can often be guided by either global coverage or explicit temporal constraints. The core difficulty lies in the third type: before answering *what* happened, the model needs to infer *where* to acquire relevant evidence.

Existing agentic systems for long-form video understanding fall into two lines. Query-agnostic pipelines apply a fixed pipeline (e.g., global scanning or precomputed textual representations) to all queries, wasting computation on irrelevant content (Yang et al., 2025a; Zuo et al., 2026; Yang et al., 2025b; He et al., 2025; Wang et al., 2026b; Yin et al., 2025; Wang et al., 2024b, 2025b; Ma et al., 2025; Luo et al., 2026; Zhang et al., 2026a). Query-conditioned pipelines enable adaptive acquisition while still relying on a single dominant strategy for inspection, particularly for implicit

* Corresponding authors.

† Project leader.

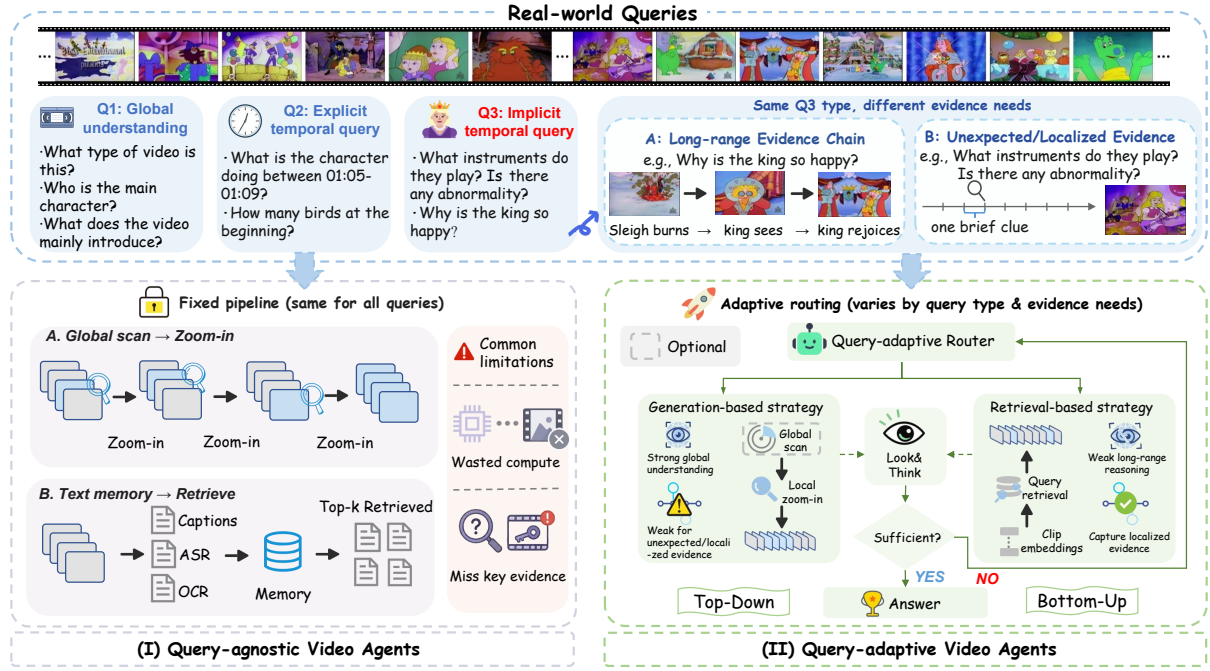


Figure 1: Real-world long-video queries demand different evidence acquisition strategies. Query-agnostic agents apply a fixed pipeline to all queries, wasting computation or missing key evidence. In contrast, Route2Look introduces query-adaptive routing that dynamically selects between top-down generation-based and bottom-up retrieval-based strategies for efficient and accurate evidence acquisition.

temporal queries. These strategies fall into two paradigms based on how temporal intervals are selected. Generation-based strategies adopt a *top-down* paradigm, reasoning over the query and accumulated observations to progressively generate temporal intervals (Lin et al., 2026a; Li et al., 2026; Zhang et al., 2026b). They excel at inference from global context or logic flow but may miss unexpected, localized events (Lin et al., 2026a). In contrast, retrieval-based strategies adopt a *bottom-up* paradigm, retrieving semantically relevant temporal intervals in parallel via clip-level embeddings (Chen et al., 2025a; Yuan et al., 2025; Chen et al., 2026). They capture salient local evidence but struggle with long-range causal chains or holistic narratives.

Overall, existing methods lack the ability to adapt evidence acquisition strategy to diverse query demands. To quantify this gap, we conduct an oracle study on a subset of LVBench (Wang et al., 2025a): if one could perfectly select between generation-based and retrieval-based strategies per query, accuracy improves from 68.5% (best single strategy) to 81.5%, yielding a substantial 13.0% absolute gain. This strongly suggests that query-adaptive evidence acquisition is essential for long-form video understanding.

Inspired by this, we propose **Route2Look**, a framework for query-adaptive evidence acquisition in long video understanding. It operates in a *Route-Look-Memorize* loop: it first routes the query to a suitable evidence acquisition strategy, then looks into the selected video evidence, and finally memorizes verified evidence for subsequent routing. The framework provides three tools: *Global Browse* for holistic context building, *Temporal Ground* for explicit temporal grounding, and *Semantic Retrieve* for parallel semantic search. To decide which tool to invoke for evidence acquisition, Route2Look learns a reusable routing skill through a two-stage design: it first distills the routing skill from differential contrastive analysis between generation-based and retrieval-based trajectories, and then applies the distilled skill with hard routing rules and continue-or-stop criteria during inference. As a result, Route2Look provides a lightweight, model-agnostic, and transferable framework for long-form video understanding.

In summary, our contributions are threefold:

- We identify query-adaptive evidence acquisition as a core challenge in long-form video understanding, demonstrating that different query types demand distinct evidence acquisition strategies instead of a single fixed one.

- We propose Route2Look, a two-stage framework that distills routing skill from differential contrastive analysis between retrieval-based and generation-based strategies, enabling query-adaptive strategy selection.
- Extensive experiments on long-form video benchmarks demonstrate that Route2Look improves both accuracy and efficiency, generalizing across datasets and query types. Oracle routing results further highlight significant potential for future query-adaptive research.

2 Related Works

2.1 Video Agentic Models

Recent advances in MLLMs have promoted agentic systems for long-form video understanding. Unlike single-pass video processing, video agents interact with videos through iterative perception, reasoning, and tool use, making them suitable for temporally extended scenarios. Existing methods can be broadly divided into two lines. The first line follows *query-agnostic pipelines*: some methods perform global scanning followed by progressive zoom-in (Yang et al., 2025a; Zuo et al., 2026; Yang et al., 2025b; He et al., 2025; Yin et al., 2025), while others precompute textual memories such as dense captions, ASR, or OCR for downstream retrieval and reasoning (Wang et al., 2024b, 2025b; Ma et al., 2025; Pang and Wang, 2025; Tian et al., 2025; Gao et al., 2025; Luo et al., 2026; Zhang et al., 2026a; Lian et al., 2026). Although effective, these methods apply a fixed processing procedure to different queries, which may waste computation on irrelevant content or miss query-specific evidence. The second line introduces *query-conditioned planning* before detailed perception (Chen et al., 2025a; Yuan et al., 2025; Wang et al., 2025c; Lin et al., 2026a; Li et al., 2026; Zhang et al., 2026b). These methods improve adaptivity by using the query to guide inspection, but they usually rely on a single dominant evidence acquisition strategy, such as global browsing with local refinement or retrieval with verification.

2.2 Automatic Skill Self-Evolution

Automatic skill self-evolution has become an important direction for improving the adaptability of language agents (Wang et al., 2023; Shinn et al., 2023; Zhao et al., 2024; Zheng et al., 2025; Yang et al., 2026b; Alzubi et al., 2026; Ni et al., 2026). Instead of relying on manually designed

prompts or fixed tool-use policies, these methods extract generalizable skills from trajectories, interactions, or execution feedback, and reuse them in future tasks. Prior work studies skill generation through exploration (Zheng et al., 2025), skill extraction from user-agent interactions (Yang et al., 2026b), structured skill memory (Jiang et al., 2026; Zhou et al., 2026), and failure-driven behavior refinement (Alzubi et al., 2026). ExpeL (Zhao et al., 2024) extracts general insights from success–failure experiences, while Trace2Skill (Ni et al., 2026) distills trajectory-local lessons into transferable declarative skills through parallel consolidation. Sharing this trajectory-to-skill paradigm, our method extends it to long-form video understanding through strategy-level differential contrastive analysis: it compares generation-based and retrieval-based trajectories for the same query–video pair using an accuracy-first and efficiency-second criterion, and distills their relative preference into a routing skill for query-adaptive evidence acquisition.

3 Method

In this work, we propose **Route2Look**, a query-adaptive framework for evidence acquisition in long-form video understanding (Fig. 2). We first formalize an agentic evidence acquisition protocol (Sec. 3.1), then present a two-stage framework that distills routing experience into reusable skill and applies it during frozen inference (Sec. 3.2).

3.1 Agentic Evidence Acquisition Protocol

Route2Look operates in a *Route-Look-Memorize* loop. Given query Q and video V of duration l , the agent iteratively acquires evidence until sufficient to answer Q . At each iteration, it selects an acquisition tool, inspects visual content, updates memory and decides whether to continue. The agent π_θ remains frozen throughout.

3.1.1 Toolkit

The agent is equipped with three evidence acquisition tools, each designed for a distinct acquisition pattern.

Global Browse. It uniformly samples frames across the full video to provide holistic context, supporting global understanding queries or providing a coarse contextual prior before more localized inspection.

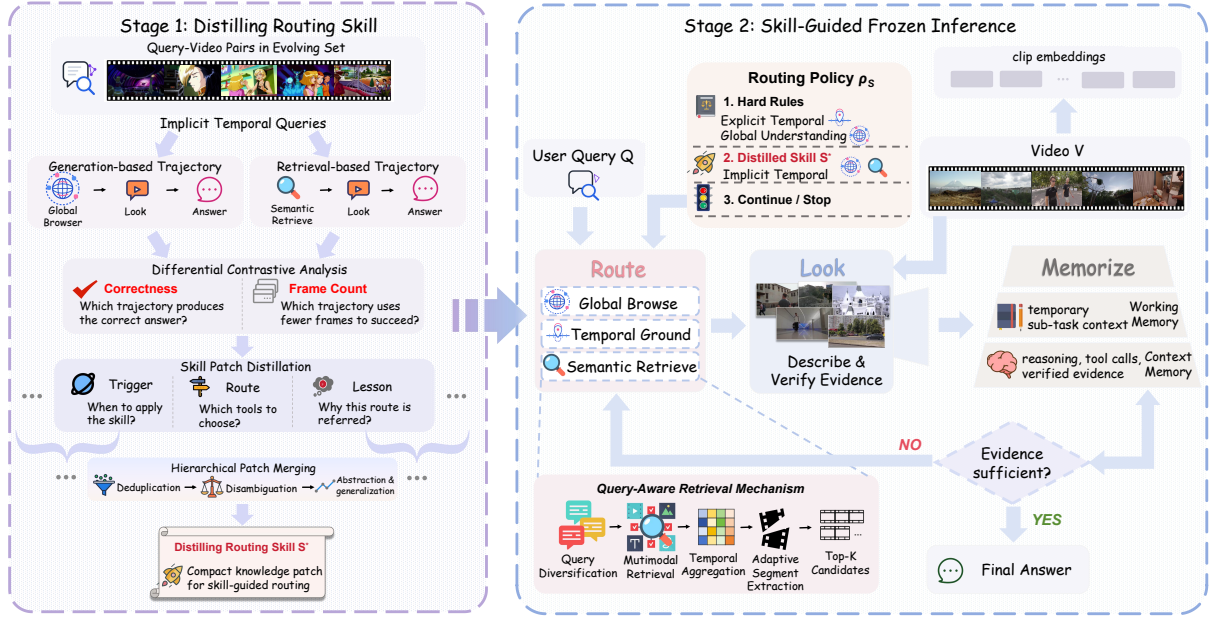


Figure 2: **Overview of Route2Look.** The agent follows a Route-Look-Memorize loop for iterative evidence acquisition. In Stage 1, Route2Look compares generation-based and retrieval-based trajectories for the same query-video pairs in an evolving set, then distills the resulting strategy preferences into a reusable routing skill. In Stage 2, the distilled routing skill is combined with hard routing rules and stopping criteria to guide query-adaptive evidence acquisition on the test set without updating the backbone model.

Temporal Ground. It uniformly samples frames within a definite temporal interval, specified by queries containing explicit timestamps, temporal ranges, or temporal descriptions.

Semantic Retrieve. It performs query-aware semantic search over the entire video, as detailed in Sec. 3.1.3. Rather than sequential browsing, it retrieves candidate clips in parallel by ranking multimodal embedding relevance to the query, followed by frame sampling from the retrieved candidates.

3.1.2 Route-Look-Memorize

Route. At iteration t , the agent receives the current query Q_t (initialized as Q and updated with memory). Given Q_t and the available tools $\mathcal{T}$, the agent performs routing reasoning r_t and selects a tool $u_t \in \mathcal{T}$ via π_θ . For global understanding queries and explicit temporal queries, the agent applies deterministic routing rules. For more challenging implicit temporal queries, the agent decides whether to adopt a generation-based strategy through Global Browse or a retrieval-based strategy through Semantic Retrieve.

Look. The selected tool returns sampled frames F_t , which are then passed to a Vision-Language Model (VLM) for detailed visual inspection. The VLM generates descriptions and assesses whether

the content provides useful evidence, filtering irrelevant segments while preserving evidence E_t for the final answer.

Memorize. Two complementary memory components are maintained: *Context memory* $\mathcal{M}_c$ stores the full interaction history $\{(r_i, u_i, E_i)\}_{i=1}^t$ to support coherent reasoning; *Working memory* $\mathcal{M}_w$ maintains focused context for sub-tasks, and only verified evidence returns to $\mathcal{M}_c$.

After each iteration, the agent evaluates whether $\mathcal{M}_c$ contains sufficient evidence to answer Q . If yes, it terminates and generates the final answer; otherwise, it proceeds to the next iteration (limited by a maximum iteration budget H).

3.1.3 Query-Aware Retrieval Mechanism

We now detail the retrieval mechanism used by Semantic Retrieve.

Query Diversification. To improve recall, we expand the current query Q_t into a set of diversified sub-queries Q_t^i which include multiple reformulations, higher-level semantic abstractions, alternative perspectives, and variants that incorporate information from multiple-choice options. This diversification increases the probability of matching relevant visual content across different semantic framings (Wang et al., 2026a).

Multimodal Retrieval. The video is divided into N_c non-overlapping clips $\{c_j\}_{j=1}^{N_c}$. Each sub-query $Q_t^i \in \mathcal{Q}_t$ and each clip c_j is encoded into a shared multimodal embedding space. We compute a temporal relevance score for each sub-query:

$$s_i[j] = \cos(\Phi(Q_t^i), \Phi(c_j)), \quad j = 1, \dots, N_c, \quad (1)$$

where Φ denotes a pre-trained multi-modal encoder, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

Temporal Evidence Aggregation. To combine signals from different query formulations, we apply element-wise max pooling over all relevance score curves:

$$s_{\text{fused}}[j] = \max_i s_i[j], \quad j = 1, \dots, N_c. \quad (2)$$

This fused score curve highlights temporal locations that match any diversified query formulation.

Adaptive Segment Extraction. Candidate segments are extracted using a dynamic threshold:

$$\gamma_{\text{ret}} = \mu(s_{\text{fused}}) + \alpha \cdot \sigma(s_{\text{fused}}), \quad (3)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation of the fused scores, and α controls the sensitivity of extraction. Peaks above γ_{ret} are identified and expanded bilaterally to cover coherent temporal events.

Candidate Selection. The extracted segments are ranked by their average fused relevance score. The top- K segments are selected as candidate evidence and passed to the Look phase for detailed verification.

3.2 Two-Stage Framework: From Experience to Skill for Query-Adaptive Routing

The evidence acquisition protocol defines how the agent acquires evidence once a tool is selected. The remaining challenge is how to choose the optimal tool for each query. To address this, Route2Look uses a two-stage framework: Stage 1 distills routing knowledge from competing trajectories on an evolving set $\mathcal{D}_{\text{evolve}}$; Stage 2 applies the distilled skill during frozen inference on test set $\mathcal{D}_{\text{test}}$.

3.2.1 Routing Skill Evolution Formalization

We define a routing skill $\mathcal{S}$ as a structured, human-readable knowledge document that specifies query-adaptive routing strategies, i.e., when to use a generation-based or retrieval-based strategy. Let

$\mathcal{D}_{\text{evolve}}$ and $\mathcal{D}_{\text{test}}$ denote two disjoint query-video sets for skill evolution and evaluation, respectively.

Given a skill $\mathcal{S}$, we measure its answer accuracy on a dataset $\mathcal{D}$ as

$$\mathcal{P}(\mathcal{S}; \pi_\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(Q, V, y^*) \in \mathcal{D}} \mathbf{1}[\pi_\theta(Q, V; \mathcal{S}) = y^*]. \quad (4)$$

where y^* denotes the ground-truth answer. We further measure frame efficiency by the average number of inspected frames:

$$\mathcal{C}(\mathcal{S}; \pi_\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(Q, V, y^*) \in \mathcal{D}} c(Q, V; \mathcal{S}), \quad (5)$$

where $c(Q, V; \mathcal{S})$ is the number of frames inspected when answering query Q on video V .

The objective of skill evolution is to construct a routing skill from competing trajectories on $\mathcal{D}_{\text{evolve}}$, without updating agent parameters θ :

$$\mathcal{P}(\mathcal{S}^*; \pi_\theta, \mathcal{D}_{\text{test}}) > \mathcal{P}(\mathcal{S}_0; \pi_\theta, \mathcal{D}_{\text{test}}), \quad (6)$$

or, when accuracies are equal,

$$\mathcal{C}(\mathcal{S}^*; \pi_\theta, \mathcal{D}_{\text{test}}) < \mathcal{C}(\mathcal{S}_0; \pi_\theta, \mathcal{D}_{\text{test}}). \quad (7)$$

The initial skill $\mathcal{S}_0$ represents a single fixed strategy (either generation-based or retrieval-based). The goal is to learn a routing skill $\mathcal{S}^*$ that outperforms the better of these two baselines.

3.2.2 Stage 1: Distilling Routing Skill from Competing Trajectories

This stage aims to attain a generalizable routing skill $\mathcal{S}^*$ from an evolving set $\mathcal{D}_{\text{evolve}}$.

Dual-Trajectory Sampling. For each query-video pair $(Q, V) \in \mathcal{D}_{\text{evolve}}$, we independently execute two distinct agentic trajectories within the pipeline defined in Sec. 3.1. For each implicit temporal query in $\mathcal{D}_{\text{evolve}}$, we execute two competing trajectories: a generation-based trajectory τ^{gen} (using Global Browse) and a retrieval-based trajectory τ^{ret} (using Semantic Retrieve). For explicit or global queries, both trajectories follow the same deterministic routing.

Differential Contrastive Analysis. A preference pair (τ^+, τ^-) is formed when one trajectory outperforms the other according to the accuracy-first, efficiency-second criterion: τ^+ is preferred if it answers correctly while τ^- fails, or if both answer correctly but τ^+ inspects fewer frames.

Table 1: Main results on three long video understanding benchmarks. #Frames denotes the number of processed frames. Best results are in **bold**, second-best are underlined.

Type	Method	LVBench		VideoMME (Long)		LongVideoBench (Long)	
		#Frames↓	Acc↑	#Frames↓	Acc↑	#Frames↓	Acc↑
VLM -based	Qwen2.5-VL-72B (Bai et al., 2025b)	768	47.3	256	53.2	-	-
	Qwen3-VL-8B (Bai et al., 2025a)	768	45.8	256	59.3	-	-
	Qwen3.5-9B (Team, 2026)	768	60.8	256	67.3	-	-
	Gemini-1.5-Pro (Team et al., 2024)	3600	33.1	1233	67.4	256	58.6
	Gemini 2.0 Flash (Pichai et al., 2024)	4037	48.3	1233	63.0	256	45.7
	GPT-4o (Hurst et al., 2024)	384	30.8	384	65.3	256	60.9
	GPT-5 (Singh et al., 2025)	384	60.1	384	67.9	384	64.5
Agent -based	VideoAgent (Wang et al., 2024b)	25.5	29.3	24.6	46.4	-	-
	VideoTree (Wang et al., 2025b)	103.2	28.8	98.0	53.1	-	-
	DrVideo (Ma et al., 2025)	-	-	493.2	51.7	-	-
	VCA (Yang et al., 2025a)	20.0	41.3	18.1	54.2	-	-
	Mr.Video (Pang and Wang, 2025)	8074	60.8	4932	61.8	2816	61.6
	DVD (Zhang et al., 2026a)	8074	<u>74.2</u>	4932	67.3	2816	68.6
	VideoSeek (Lin et al., 2026a)	92.3	68.4	60.9	<u>70.1</u>	29.6	<u>73.5</u>
	Route2Look (Ours)	202.3	75.4	126.6	76.1	168.7	77.8

Skill Patch Distillation. For each preference pair, π_θ distills a skill patch $p_i = (Trigger_i, Strategy_i, Lesson_i)$, where *Trigger* describes the query characteristics, *Strategy* recommends either generation or retrieval, and *Lesson* provides actionable routing guidelines. The patch converts instance-level trajectory comparisons into reusable routing knowledge.

Hierarchical Patch Merging. The extracted patches may be redundant or conflicting. We merge them hierarchically: patches are grouped into batches and iteratively merged via the LLM π_θ , which removes duplicates, resolves conflicts by favoring higher-support patches, and abstracts instance-specific observations into general principles. This process repeats until the number of patches falls below the batch size, yielding a compact routing skill $\mathcal{S}^*$.

3.2.3 Stage 2: Skill-Guided Frozen Inference

During inference, the distilled skill $\mathcal{S}^*$ is injected into the agent’s routing policy $\rho_{\mathcal{S}}$, which consists of three layers. **Layer 1** applies hard routing rules: Temporal Ground for explicit temporal hints and Global Browse for global understanding queries. **Layer 2** employs the learned skill $\mathcal{S}^*$ to guide implicit temporal queries toward either retrieval-based or generation-based strategies. **Layer 3** implements continue-or-stop criteria, terminating the loop when sufficient evidence is accumulated or the maximum iteration budget H is reached.

4 Experiments

4.1 Experimental Settings

Benchmarks. We evaluate Route2Look on three long-form video understanding benchmarks. LVBench (Wang et al., 2025a) contains 1,549 multiple-choice questions over 103 hour-long videos. VideoMME (Fu et al., 2025) (long subset) has 900 questions with an average duration of 2,466 seconds. LongVideoBench (Wu et al., 2024) (long split of its validation set) includes 564 questions from 188 videos lasting 900–3,600 seconds. Together, these benchmarks cover a broad spectrum of long-form video scenarios, enabling a comprehensive evaluation of both the accuracy and efficiency of query-adaptive evidence acquisition.

Baselines. We compare against two categories of methods for video understanding: (1) VLM-based methods, covering both commercial and open-source models, such as GPT-5, GPT-4o, Gemini-1.5 Pro, Gemini 2.0 Flash, and Qwen2.5-VL-72B. (2) Agent-based methods, such as VideoAgent, Videotree, VCA, MR.Video, DVD, VideoSeek which also use GPT-5 as the backbone. Most reported results are from official leaderboards or published reports.

Implementation Details. Route2Look is model-agnostic and can be instantiated with different LLM/VLM backbones. By default, we use GPT-5 as the agent backbone and LanguageBind (Zhu et al., 2024) as the multimodal retriever using 10-second clips. We set top- K to 5, the threshold parameter α to 0.5, and the maximum number of

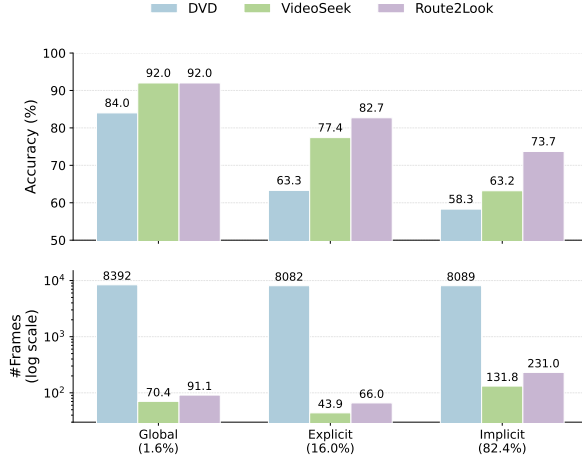


Figure 3: Performance breakdown by query type on LVBench. The upper panel shows accuracy, and the lower panel shows frame usage in log scale. Query-type ratios are shown in parentheses on the x-axis.

agent iterations to 10. For each tool call, videos are sampled at 2 FPS with at most 50 frames due to API input limits. The evolving set contains 500 stratified samples from CG-Bench (Chen et al., 2025b), and the merge batch size for hierarchical patch merging is set to $B = 32$. Hyperparameter analysis is provided in Appendix A.1.

4.2 Main Results

Table 1 shows that Route2Look achieves state-of-the-art performance on all benchmarks while maintaining comparable frame efficiency. On LVBench, it reaches 75.4% accuracy with 202.3 frames, outperforming DVD by 1.2% while using only 2.5% of its frame budget. On VideoMME, Route2Look achieves 76.1% with 126.6 frames, surpassing VideoSeek by 6.0%. On LongVideoBench, it obtains 77.8% with 168.7 frames, improving over DVD and VideoSeek by 9.2% and 4.3%, respectively. These consistent gains are achieved with the routing skill distilled from an external evolution set, showing that Route2Look learns generalizable routing patterns and improves the base model GPT-5 with substantially lower frame usage.

4.3 Performance on Different Query Types

We further analyze performance by query type in Fig. 3, including *global understanding*, *explicit temporal*, and *implicit temporal* queries. We reproduce two strong baselines, DVD and VideoSeek, with GPT-5 as the same backbone for a fair comparison. The three query types account for 1.6%, 16.0%, and 82.4% of LVBench, respectively, show-

Table 2: Ablation studies on core components on LVBench. “Gen.” and “Ret.” denote generation-based and retrieval-based strategies; “Diff. Con.” denotes differential contrastive analysis.

Gen.	Ret.	Routing	Diff. Con.	Acc	#Frames
✓		–	–	63.5	107.5
	✓	–	–	68.5	290.7
✓	✓	Random	–	65.8	196.8
✓	✓	Heuristic	–	66.5	255.4
✓	✓	Learned	✗	61.5	238.6
✓	✓	Learned	✓	70.0	244.9
✓	✓	Oracle	–	81.5	131.8

ing that implicit temporal queries dominate the benchmark. Route2Look achieves 92.0% accuracy on global understanding queries, matching VideoSeek and surpassing DVD by 8.0%. For explicit temporal queries, it improves over VideoSeek from 77.4% to 82.7% with only 66.0 frames on average. The largest gain appears on implicit temporal queries, where Route2Look reaches 73.7% accuracy, outperforming VideoSeek by 10.5% and DVD by 15.4%. Overall, Route2Look achieves the best accuracy with competitive frame efficiency across query types, especially on the dominant and challenging implicit temporal queries.

4.4 Ablation Studies

We conduct ablations on a 200-sample stratified subset of LVBench that preserves the domain composition and accuracy trends of the full dataset. More details are provided in Appendix A.4.

Effectiveness of Core Components. Since global and explicit temporal queries are handled by deterministic rules, this ablation focuses on implicit temporal queries where routing is most critical. We compare Route2Look with Generation-only, Retrieval-only, Random routing, Heuristic routing, w/o differential contrastive analysis, and Oracle. As for Heuristic routing, we implement a non-learned keyword-based router that uses explicit surface cues in the query. For Heuristic routing, we implement a non-learned keyword-based router using explicit surface cues in the query. Cues such as “why,” “relationship,” and “throughout the video” are routed to Global Browse, while cues such as “what color,” “how many,” and “what is written” are routed to Semantic Retrieve. The w/o differential contrastive analysis variant summarizes each trajectory independently, while Oracle selects

Browse or Temporal Ground and obtain correct answers with as few as 50 frames. For implicit temporal queries, Route2Look adapts its strategy to the evidence need: it uses Semantic Retrieve for localized evidence, Global Browse for persistent global attributes, and multiple tools for complex movement reasoning. These cases show that implicit queries do not favor a fixed strategy, and Route2Look can flexibly route among strategies across different queries and within a single query.

5 Conclusion

We present Route2Look, a lightweight and model-agnostic framework for query-adaptive evidence acquisition in long-form video understanding. Motivated by the mismatch between query demands and fixed evidence acquisition strategies, Route2Look routes before looking and dynamically selects among three evidence acquisition tools. It learns reusable routing skills from contrastive generation-based and retrieval-based trajectories, enabling adaptive inference without updating the backbone model. Experiments on three long-video benchmarks demonstrate state-of-the-art accuracy with comparable frame efficiency. Further analyses show that the gains are most pronounced on implicit temporal queries, confirming the importance of adapting evidence acquisition to query-specific evidence needs.

Limitations

Route2Look provides a first step toward query-adaptive evidence acquisition, and several limitations remain. First, the routing skill is distilled from a finite evolution set, leaving room to close the gap between learned and oracle routing through larger data or better skill optimization. Second, Route2Look still requires moderate frame verification, especially when Semantic Retrieve returns multiple candidates. Future work can explore more selective verification policies to further reduce frame usage while maintaining or improving accuracy. Third, the current framework mainly focuses on visual evidence, while audio cues may provide complementary signals for long-form video understanding. These limitations together indicate broad opportunities for future adaptive routing research. By revealing the importance and room of query-adaptive strategy selection, Route2Look lays a foundation for building more accurate, efficient, and multimodally grounded long-video agents.

Acknowledgements

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0123404) and the Beijing Natural Science Foundation (No. L252035).

References

- Salaheddin Alzubi, Noah Provenzano, Jaydon Bingham, Weiyuan Chen, and Tu Vu. 2026. Evoskill: Automated skill discovery for multi-agent systems. *arXiv preprint arXiv:2603.02766*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint, arXiv:2502.13923*.
- Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. 2025a. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20237–20246.
- Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. 2025b. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pages 45647–45682.
- Tao Chen, Kun Zhang, Qiong Wu, Xiao Chen, Chao Chang, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. 2026. Scaling the long video understanding of multimodal large language models via visual memory mechanism. *arXiv preprint arXiv:2603.29252*.
- Chaoyou Fu, Yuhao Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.
- Hong Gao, Yiming Bao, Xuezhen Tu, Yutong Xu, Yue Jin, Yiyang Mu, Bin Zhong, Linan Yue, and Min-Ling Zhang. 2025. Agentic video intelligence: A flexible framework for advanced video exploration and understanding. *arXiv preprint arXiv:2511.14446*.

- Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. 2025. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Guanyu Jiang, Zhaochen Su, Xiaoye Qu, and Yi R Fung. 2026. Xskill: Continual learning from experience and skills in multimodal agents. *arXiv preprint arXiv:2603.12056*.
- Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, et al. 2025. A survey on vision-language-action models for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4524–4536.
- Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. 2026. Lenswalk: Agentic video understanding by planning how you see in videos. *arXiv preprint arXiv:2603.24558*.
- Niu Lian, Yuting Wang, Hanshu Yao, Jinpeng Wang, Bin Chen, Yaowei Wang, Min Zhang, and Shu-Tao Xia. 2026. From verbatim to gist: Distilling pyramidal multimodal memory via semantic information bottleneck for long-horizon video agents. *arXiv preprint arXiv:2603.01455*.
- Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. 2026a. Videoseek: Long-horizon video agent with tool-guided seeking. *arXiv preprint arXiv:2603.20185*.
- Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, et al. 2026b. Unleashing hour-scale video training for long video-language understanding. *Advances in Neural Information Processing Systems*, 38:17523–17552.
- Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. 2026. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033.
- Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezaatofghi, and Jianfei Cai. 2025. Dvideo: Document retrieval based long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18936–18946.
- Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjuan Jiang. 2026. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158*.
- Ziqi Pang and Yu-Xiong Wang. 2025. Mr. video: "mapreduce" is the principle for long video understanding. *arXiv preprint arXiv:2504.16082*.
- Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2024. Introducing gemini 2.0: Our new ai model for the agentic era. <https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/>. Google Blog, accessed May 16, 2026.
- Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24045–24055.
- Chenhui Qiang, Zhaoyang Wei, Xumeng Han, Zipeng Wang, Siyao Li, Xiangyuan Lan, Jianbin Jiao, and Zhenjun Han. 2025. Ver-bench: Evaluating mllms on reasoning with fine-grained visual evidence. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12698–12705.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.
- Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.
- Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. 2025. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Qwen Team. 2026. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.

- Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkan Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*.
- Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Tianyue Wang, Leigang Qu, Tianyu Yang, Xiangzhao Hao, Yifan Xu, Haiyun Guo, and Jinqiao Wang. 2026a. Wiser: Wider search, deeper thinking, and adaptive fusion for training-free zero-shot composed image retrieval. *arXiv preprint arXiv:2602.23029*.
- Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. 2025a. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.
- Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.
- Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. 2026b. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. *arXiv preprint arXiv:2603.04977*.
- Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025b. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283.
- Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. 2025c. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*.
- Keshara Weerasinghe, Saahith Janapati, Xueren Ge, Sion Kim, Sneha Iyer, John A Stankovic, and Homa Alemzadeh. 2024. Real-time multimodal cognitive assistant for emergency medical services. *arXiv preprint arXiv:2403.06734*.
- Zhaoyang Wei, Chenhui Qiang, Bowen Jiang, Xumeng Han, Xuehui Yu, and Zhenjun Han. 2025. Ad² bench: A hierarchical cot benchmark for mllm in autonomous driving under adverse conditions. *arXiv preprint arXiv:2506.09557*.
- Zhaoyang Wei, Guoliang Wang, Guohua Gao, Yanchao Hao, Mingda Li, Wenchao Ding, Xi Chen, Shizhu He, and Xuehui Yu. 2026. Seeing is believing: Grounding long-video understanding in spatio-temporal visual evidence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10584–10592.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.
- Haiyang Yan, Hongyun Zhou, Peng Xu, Xiaoxue Feng, and Mengyi Liu. 2026. Symphony: A cognitively-inspired multi-agent system for long-video understanding. *arXiv preprint arXiv:2603.17307*.
- Ruoliang Yang, Chu Wu, Caifeng Shan, Ran He, and Chaoyou Fu. 2026a. Videodetective: Clue hunting via both extrinsic query and intrinsic relevance for long video understanding. *arXiv preprint arXiv:2603.22285*.
- Yutao Yang, Junsong Li, Qianjun Pan, Bihao Zhan, Yuxuan Cai, Lin Du, Jie Zhou, Kai Chen, Qin Chen, Xin Li, et al. 2026b. Autoskill: Experience-driven life-long learning via skill self-evolution. *arXiv preprint arXiv:2603.01145*.
- Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2025a. Vca: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20168–20179.
- Zhongyu Yang, Zuhao Yang, Shuo Zhan, Tan Yue, Wei Pang, and Yingfang Yuan. 2026c. Svagent: Storyline-guided long video understanding via cross-modal multi-agent collaboration. *arXiv preprint arXiv:2604.05079*.
- Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. 2025b. Longvt: Incentivizing "thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*.
- Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. 2025. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. *arXiv preprint arXiv:2512.12360*.

- Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. 2025. Videoexplorer: Think with videos for agentic long-video understanding. *arXiv preprint arXiv:2506.10821*.
- Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. 2026a. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895.
- Yaolun Zhang, Ruohui Wang, Jiahao Wang, Yepeng Tang, Xuanyu Zheng, Haonan Duan, Hao Lu, Hanming Deng, and Lewei Lu. 2026b. Eva: Efficient reinforcement learning for end-to-end video agent. *arXiv preprint arXiv:2603.22918*.
- Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.
- Shitian Zhao, Shaoheng Lin, Ming Li, Haoquan Zhang, Wenshuo Peng, Kaipeng Zhang, and Chen Wei. 2026. Pyvision-rl: Forging open agentic vision models via rl. *arXiv preprint arXiv:2602.20739*.
- Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. 2025. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*.
- Huichi Zhou, Siyuan Guo, Anjie Liu, Zhongwei Yu, Ziqin Gong, Bowen Zhao, Zhixun Chen, Menglong Zhang, Yihang Chen, Jinsong Li, et al. 2026. Memento-skills: Let agents design agents. *arXiv preprint arXiv:2603.18743*.
- Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, volume 2024, pages 9588–9608.
- Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.
- Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkan Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. 2026. Videolucy: Deep memory backtracking for long video understanding. *Advances in Neural Information Processing Systems*, 38:25660–25691.

A Appendix

A.1 Hyperparameter Analysis

We analyze three key hyperparameters in Table 4: the evolution set size $|\mathcal{D}_{\text{evolve}}|$, the merge batch size B , and the number of retrieved segments K .

Evolution set size. Increasing the evolution set from 200 to 500 samples yields a notable improvement from 67.0% to 70.0%, indicating that more diverse training queries help distill a more robust routing skill. This trend suggests that further scaling the evolution data may continue to improve routing quality.

Merge batch size. With 200 samples, batch size 32 achieves the best accuracy (67.0%), while extremely small and overly large batches degrade performance. A similar trend holds with 500 samples, where increasing batch size from 32 to 200 causes a drop from 70.0% to 64.0%. These results indicate that merging too many patches at once loses fine-grained distinctions, while too few patches per batch fail to resolve conflicts effectively. A moderate batch size $B = 32$ consistently yields the best trade-off.

Number of retrieved segments. Decreasing K from 5 to 3 substantially reduces frame usage from 249.5 to 171.7, but at the cost of a notable accuracy drop from 67.0% to 61.5%. This indicates that retrieving fewer candidates improves efficiency but risks missing relevant evidence, and $K = 5$ provides a favorable balance between coverage and cost.

Table 4: Hyperparameter analysis on LVBench.

$ \mathcal{D}_{\text{evolve}} $	B	K	Acc	#Frames
200	32	3	61.5	171.7
200	32	5	67.0	249.5
200	10	5	63.5	186.0
200	200	5	66.5	216.8
500	32	5	70.0	244.9
500	200	5	64.0	253.7

A.2 Efficiency Analysis

We further analyze the efficiency of Route2Look on LVBench. We report the average frame usage and token consumption of the GPT-5 base model and Route2Look in Table 5. Compared with the GPT-5 base model, Route2Look uses substantially

Table 5: Performance comparison of different methods.

Method	#Frames	#Tokens	Acc
GPT-5 (Base)	384	83K	60.1
Route2Look (Ours)	202.3	72K	75.4

fewer frames and fewer tokens, while achieving much higher accuracy. Specifically, Route2Look reduces the average number of inspected frames from 384.0 to 202.3, corresponding to a 47.3% reduction, and decreases token consumption from 83K to 72K. Meanwhile, the accuracy improves from 60.1% to 75.4%. These results demonstrate that query-adaptive evidence acquisition enables the agent to focus on more relevant video segments, leading to a better accuracy-efficiency trade-off.

A.3 Algorithm Demonstration

Algorithm 1 illustrates the agent workflow in Route2Look. The entire process follows a Route-Look-Memorize loop. Given a query Q and a video V , the agent repeatedly performs three steps. First, it updates the current query state Q_t based on context memory $\mathcal{M}_c$ and selects a tool u_t from the toolkit $\mathcal{T}$ under the routing policy ρ_S . Second, the agent initializes a temporary working memory $\mathcal{M}_w^t$ for the current tool call. The selected tool is then invoked to obtain sampled frames F_t , and the VLM inspects these frames to extract verified evidence E_t . Third, only the verified evidence and routing information are written back to the context memory. And the agent checks whether the accumulated evidence is sufficient to answer the query. If so, it generates the final answer Y ; otherwise, it continues the loop until reaching the maximum iteration budget H .

A.4 Subset Analysis in the Ablation Study

We conduct ablation studies on a stratified subset of 200 samples from LVBench to reduce evaluation cost while preserving the benchmark distribution. Following the sampling protocol of (Yin et al., 2025), we sample from the joint domain-task distribution of the full set, prioritizing domain-level proportions and maintaining task-level proportions within each domain. Samples are drawn without replacement within each stratum, with minor rounding adjustment to satisfy the 200-sample budget.

As shown in Fig. 5, the resulting subset closely matches the domain composition of the full LVBench set and preserves similar accuracy trends

Algorithm 1 Route2Look Agent Workflow

Require: User query Q ; video V ; routing policy ρ_S ; frozen LLM/VLM agent π_θ ; toolkit $\mathcal{T}$; maximum iteration budget H .

Ensure: Answer Y .

```

1: Initialize context memory  $\mathcal{M}_c \leftarrow \emptyset$ 
2: Initialize final answer  $Y \leftarrow \emptyset$ 
3: for  $t = 1$  to  $H$  do
4:   Initialize working memory  $\mathcal{M}_w^t \leftarrow \emptyset$ 
5:    $(r_t, u_t, Q_t) \leftarrow \pi_\theta(Q, \mathcal{M}_c, \mathcal{T}, \rho_S)$   $\triangleright$ 
     Routing reasoning, tool selection, the current
     query
6:    $\mathcal{M}_w^t \leftarrow \mathcal{M}_w^t \cup \{Q_t\}$   $\triangleright$  Temporary
     memory for the current tool call
7:    $(F_t, \mathcal{M}_w^t) \leftarrow \text{APPLYTOOL}(u_t, V, \mathcal{M}_w^t)$   $\triangleright$ 
     Acquire sampled frames
8:    $(E_t, \mathcal{M}_w^t) \leftarrow \pi_\theta(F_t, \mathcal{M}_w^t)$   $\triangleright$  Inspect
     frames and verify evidence
9:    $\mathcal{M}_c \leftarrow \mathcal{M}_c \cup \{r_t, u_t, E_t\}$   $\triangleright$  Update
     context memory
10:  if  $\text{SUFFICIENT}(Q, \mathcal{M}_c)$  then
11:     $Y \leftarrow \text{ANSWER}(Q, \mathcal{M}_c)$ 
12:    break
13:  end if
14: end for
15: if  $Y = \emptyset$  then
16:    $Y \leftarrow \text{ANSWER}(Q, \mathcal{M}_c)$   $\triangleright$  Forced final
     answer
17: end if
18: return  $Y$ 

```

across domains. This indicates that the subset provides a representative and efficient testbed for controlled ablation analysis.

A.5 Oracle Routing Analysis

We conduct an in-depth analysis of the oracle routing results, as summarized in Fig. 6. The results show that 50.5% of samples can be correctly answered by both strategies, indicating that a substantial portion of implicit temporal queries can be handled reliably by either top-down reasoning or bottom-up semantic retrieval. Generation-only succeeds on 13.0% of samples where retrieval fails, while retrieval-only succeeds on 18.0% of samples where generation fails. By selecting the correct strategy for each sample, oracle routing achieves 81.5% accuracy, which corresponds to the union of the two single-strategy success sets. This is substantially higher than either generation-only (63.5%) or retrieval-only (68.5%) performance, revealing

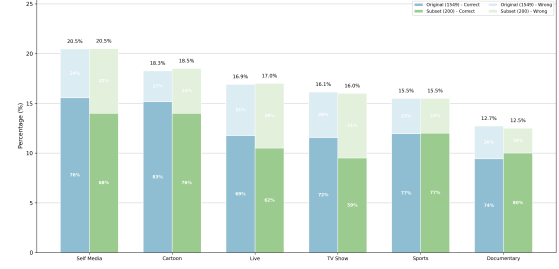


Figure 5: Effectiveness of Subset Analysis.

substantial room for query-adaptive routing. Overall, Fig. 6 provides direct empirical evidence that adaptive routing is necessary: the key challenge is not to replace one strategy with another, but to learn when each strategy should be trusted.

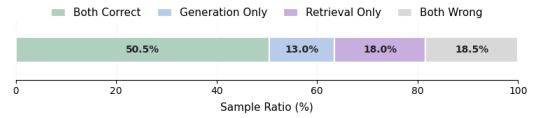


Figure 6: Oracle Routing Analysis.

A.6 Generalization Beyond the Evolving Set

Since the routing skill is distilled from a finite evolving set, an important question is whether it generalizes beyond the observed data. We examine this question from three perspectives: cross-benchmark transfer, held-out-domain transfer, and the effect of evolving-set scale.

Cross-Benchmark Generalization. The routing skill is distilled exclusively from CG-Bench and directly applied to LVBench, VideoMME, and LongVideoBench without re-distillation or benchmark-specific adaptation. Its consistent gains across these independent benchmarks indicate that the learned skill captures transferable query-evidence acquisition patterns rather than dataset-specific shortcuts.

Held-out-Domain Evaluation. To more directly assess domain generalization, we conduct held-out-domain experiments on two representative video domains, as shown in Table 6. For each target domain, we remove all corresponding samples from the CG-Bench evolving set, distill the routing skill using only the remaining domains, and evaluate it on the corresponding LVBench domain. Specifically, we evaluate transfer from CG-Bench excluding *Sports & Exercise* to the LVBench *Sport*

Table 6: Held-out-domain generalization on LVBench. The target domain is excluded from the CG-Bench evolving set during skill distillation.

Domain	Gen.-only	Ret.-only	Held-out Skill
Sport	71.0	61.3	74.2
TV	56.2	56.2	59.4

domain, and from CG-Bench excluding *Music & TV Show* to the LVBench *TV* domain.

Despite never observing samples from the corresponding domains during skill distillation, the held-out skill outperforms both generation-only and retrieval-only baselines, demonstrating its ability to transfer to unseen video domains.

Effect of Evolving-Set Scale. Increasing the evolving set from 200 to 500 samples improves accuracy from 67.0% to 70.0% in Table 4, indicating that broader coverage further improves routing quality.

A.7 Visualization of Skill

To better understand what Route2Look learns during skill evolution, we visualize part of the distilled routing skill in Table 7. The distilled skills reveal clear routing preferences. Generation-based strategy are generally favored when the required evidence can be obtained from global context, stable visual attributes, contiguous workflows, or visually clear temporal boundaries. For example, queries about overall video type, persistent props, structured processes, or visually defined durations often benefit from first building a coarse global timeline and then performing light local verification. In contrast, retrieval-based strategy are preferred when the evidence is precise, sparse, transient, or lexically grounded. Examples include exact on-screen text, brief UI changes, localized spatial relations, sparse first occurrences, or repeated similar actions that require high-recall candidate search and local verification. These examples show that the learned skill is not a simple preference for one fixed strategy. Instead, it captures reusable routing principles that connect query characteristics with evidence acquisition behavior. This human-readable form also makes the routing policy interpretable: Route2Look learns when to rely on global reasoning and when to retrieve localized evidence.

A.8 Additional Case Studies

We provide additional qualitative examples in Fig. 7–Fig. 11 to further illustrate how Route2Look adapts evidence acquisition to different query types and evidence needs. For global understanding queries in Fig. 7, Route2Look directly invokes Global Browse to capture whole-video context, such as video category, documentary topic, and overall style. For explicit temporal queries in Fig. 8, it applies Temporal Ground to inspect the specified timestamp, time range, or temporal description, enabling direct verification of localized temporal evidence.

For implicit temporal queries, Route2Look exhibits more diverse routing behaviors. When the query requires holistic judgment, repeated-event counting, global ordering, or long-range reasoning, as shown in Fig. 9, Route2Look uses Global Browse to aggregate evidence across the full video. When the answer depends on localized visual cues, such as a specific reaction, object state, or existence of a target event, it uses Semantic Retrieve to locate relevant temporal segments before verification, as shown in Fig. 10. In more ambiguous cases in Fig. 11, Route2Look can switch between tools: it first attempts Semantic Retrieve for targeted evidence and then falls back to Global Browse or performs Semantic Retrieve when the initial Global Browse is insufficient. These cases further show that implicit temporal queries do not favor a fixed strategy, and that Route2Look can adaptively route both across different queries and within a single query.



Figure 7: Additional case studies on global understanding queries. Route2Look invokes Global Browse to capture whole-video context and answer questions about video category, documentary topic, and overall style.

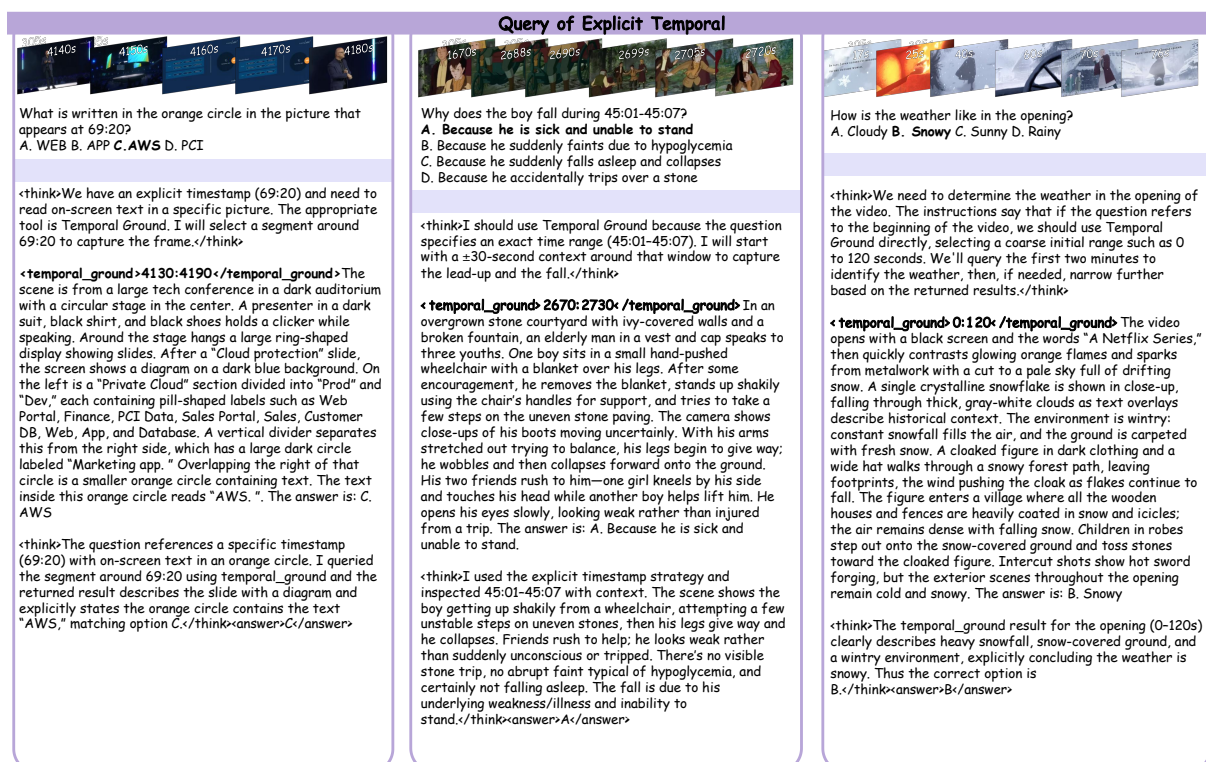


Figure 8: Additional case studies on explicit temporal queries. Route2Look invokes Temporal Ground when the query specifies a timestamp, time range, or temporal segment, enabling direct verification of localized evidence.



Figure 9: Additional case studies where implicit temporal queries favor Global Browse. When the query requires repeated-event counting, global ordering, or long-range reasoning, Route2Look uses Global Browse to aggregate evidence across the full video.



Figure 10: Additional case studies where implicit temporal queries favor Semantic Retrieve. When the answer depends on localized visual cues, such as a specific reaction, a brief object state, or the existence of a target event, Route2Look uses Semantic Retrieve to locate relevant segments before verification.



Figure 11: Additional case studies showing adaptive routing within implicit temporal queries. Route2Look can switch between Semantic Retrieve and Global Browse when the initial route is insufficient, showing adaptive routing within a single query.

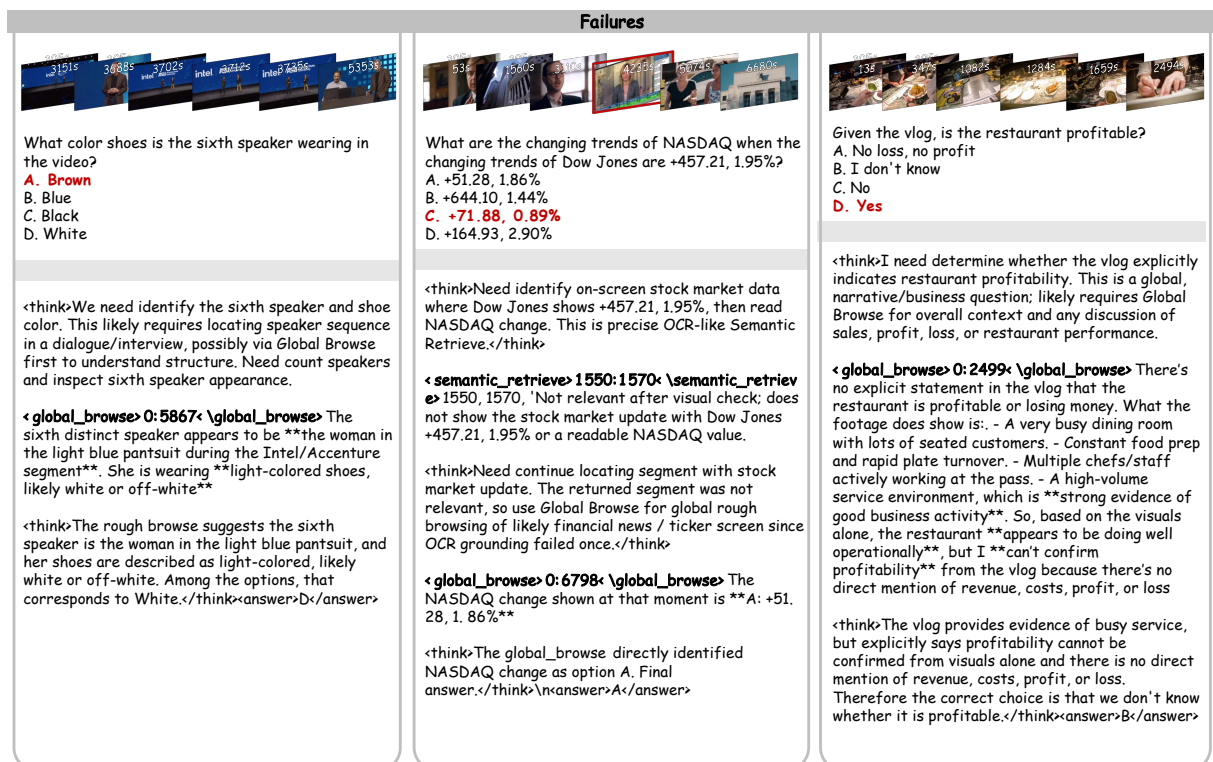


Figure 12: Representative failure cases. Route2Look may still fail on subtle spatial details, OCR-like numerical reading, or questions requiring reasoning beyond directly observable visual evidence.

Table 7: Examples of distilled routing skill.

Trigger	Route	Lesson
Global or persistent visual understanding, such as overall setting, participants, storyline flow, stable scene layout, persistent props, HUD/UI semantics, or category and role identification.	Generation-based	Skim the whole video to map scenes, participants, timeline, and causal flow. Use sparse uniform sampling to confirm stable cues and answer from representative frames. Escalate only when multiple scenes or configurations create ambiguity.
Exact linguistic strings, including on-screen text, subtitles, spoken dialogue, earliest textual instances, or narration-driven entity introductions.	Retrieval-based	Use retrieval to jump to the exact segment. Densely sample adjacent frames, crop or zoom when needed, and verify exact strings across multiple frames. For earliest-instance questions, confirm that no earlier true mention exists.
Linguistic content where exact wording is not critical, such as prominent on-screen branding, stable text layout, or a single spoken fact concentrated in one short segment.	Generation-based	Skim to capture a clear card or overlay and lightly ground the timing. Read a few clean frames or a tight transcript window. Switch to retrieval when the text is small, transient, ambiguous, or requires precise lexical distinction.
Event-anchored localized questions tied to a clear temporal cue, such as when, after, before, beginning, first, last, explicit timestamps, or quoted lines.	Generation-based	Locate the anchor through visual cues, or transcript keywords, then bound a narrow temporal window and sample adjacent frames. Verify the simple detail, immediate outcome, or directionality without heavy retrieve-verify loops.
Precise, localized, and time-bound visual evidence, including fine-grained spatial relations, transient UI labels, brief color or state changes, and multi-option frame-level cues.	Retrieval-based	Anchor the target moment through semantic or event-based retrieval, then densely sample a tight frame window. Inspect the relevant region, resolve viewpoint changes, track entities across cuts, and verify each claim against explicit evidence.
Counting repeated but visually salient items or actions within a clearly bounded interval or a stable wide shot.	Generation-based	Establish the interval boundary with a global skim, then tally within the contiguous window. Use adjacent frames to stabilize visibility, broaden slightly when needed, and track objects to avoid double-counting.
Counting or enumeration tied to a specific narrative moment, step, brief visibility window, or ordinal relation.	Retrieval-based	Retrieve the correct segment using the narrative cue or event boundary, then perform dense local verification across nearby frames. Count each visible item carefully and avoid conflating it with similar steps.
Counting how many times a repeated, visually similar action occurs across the whole video, especially with replays, multi-angle cuts, or montages.	Retrieval-based	Retrieve candidate segments across the timeline. Verify strict start and end boundaries locally, then deduplicate cross-angle replays, montage repeats, and near-duplicate views before counting.
Order-of-events reasoning in a structured workflow, tutorial, or contiguous process, including identifying the earliest or latest acted-on item.	Generation-based	Build a coarse global timeline through a quick skim, pin salient anchor actions, and read forward through the contiguous segment. Use brief local checks only when similar candidates remain.
Fine-grained temporal ordering with visually similar candidates, strong distractors, or subtle action boundaries.	Retrieval-based	Anchor the referenced cue through transcript, UI, or visual semantics. Densely sample the immediate window, track entities across cuts, and verify timestamps or boundary states until the order is resolved.
Earliest or first occurrence queries across the whole video when the target is reasonably frequent and the video is not long or cluttered.	Generation-based	Perform a global sweep to find the earliest relevant entry point, then ground and confirm the first true action or event in a tight post-cue window. Switch to retrieval if the search becomes exhaustive.
Earliest or first occurrence in long or cluttered videos where the target event is brief, sparse, or difficult to find with high recall.	Retrieval-based	Use semantic retrieval to propose candidate moments, perform dense local verification, deduplicate near-misses, and back-check earlier windows to ensure the first true instance.
Duration questions with visually defined boundaries, such as a continuous action, the n -th occurrence in a known segment, or a visually distinctive activity interval.	Generation-based	Use a global sweep to locate clear start and end boundaries, then compute elapsed time or match the closest option. Avoid sparse semantic retrieval that may miss boundary frames.