# ROUTE2LEARN: NATIVE TOOL ROUTING WITH A SINGLE MLLM FOR LONG VIDEO UNDERSTANDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long-form video understanding requires effective evidence acquisition strategies, as relevant information is often sparsely distributed across extended temporal spans. Different queries can benefit from different strategies: generation-based strategy excels at reasoning over global context, while retrieval-based strategy is better suited for locating specific events. However, existing methods either rely on a single strategy or decouple evidence acquisition from policy optimization via external modules, limiting their adaptability. In this paper, we propose **Route2Learn**, an end-to-end native agentic framework that learns query-adaptive routing between generation-based and retrieval-based strategies within a single video MLLM. Building on the insight that *learning to route requires comparing both routes*, we introduce a two-stage optimization approach: For Supervised Fine-Tuning, we actively execute both strategies for each query and progressively collect trajectories with rejection sampling, providing an implicit routing signal by exposing the model to which strategies succeed on different queries. We then introduce Counterfactual Routing Policy Optimization (CRPO), which explicitly optimizes routing preferences using the outcome difference between factual and counterfactual rollouts, while decoupling advantage estimation for tool selection and answer generation to enable precise credit assignment. Extensive experiments demonstrate that Route2Learn substantially outperforms existing strong baselines across four long-form video understanding benchmarks, improving over Qwen3-VL-8B by +10.4% on average. Further analysis shows that the model learns to select the more effective evidence-acquisition strategy according to the query. Our data, code, and model weights will be released publicly.

## 1 INTRODUCTION

Long-form video understanding is an essential capability for question answering over extended real-world videos, requiring effective evidence acquisition. However, relevant visual evidence is often sparsely distributed across long temporal spans, which makes it difficult to identify the information needed for accurate answers (Wang et al., 2024; Ye et al., 2025; Li et al., 2026b; Wang et al., 2026e).

Recent work has shown that different queries often require different evidence acquisition strategies (Wang et al., 2026c): some benefit from *generation-based strategy* that reasons over the query and accumulated observations to progressively generate temporal intervals, while others benefit from *retrieval-based strategy* that retrieves semantically relevant frames for local evidence. For effective evidence acquisition, Route2Look (Wang et al., 2026c) follows a route-before-look framework to first select an evidence acquisition strategy and then inspect the corresponding evidence. Moreover, its oracle study demonstrates the substantial benefit of adaptive routing between generation-based and retrieval-based strategies. However, it remains a training-free framework that relies heavily on instruction following to elicit the intrinsic understanding capabilities of the model, causing high sensitivity to prompt variations.

Inspired by the success of the "Think with Image" paradigm (Su et al., 2026; Wu et al., 2026), recent works have extended native tool-calling capabilities to video, giving rise to the "Think with Video" paradigm, which mainly falls into two lines. The first line enables multimodal large language models (MLLMs) to perform only generation-based strategy by dynamically selecting and re-inspecting relevant video segments (Yang et al., 2026a;b; Zeng et al., 2026; Zhang et al., 2026b). They excel



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.

at reasoning over global context or logical flow but may miss unexpected, localized events. The second line incorporates retrieval-based strategy through an auxiliary model while optimizing only the reasoning model (Yuan et al., 2025; Li et al., 2026a; Qiu et al., 2026). This reliance on external modules decouples evidence acquisition from policy optimization, limiting the potential for joint adaptation. Together, these limitations motivate our central question: *Can a single MLLM jointly support generation-based strategy and retrieval-based strategy via a dynamic routing policy?*

To answer this question, we propose **Route2Learn**, an end-to-end native agentic framework that learns query-adaptive routing between generation-based and retrieval-based strategies within a single model for long-form video understanding. Route2Learn equips a single video MLLM with two complementary tools: a generation-based tool that generates temporal intervals for targeted inspection and a retrieval-based tool that retrieves relevant frames using representations from the MLLM. Given a query, the model learns to route to the appropriate tool and then iteratively gathers visual evidence before generating the answer.

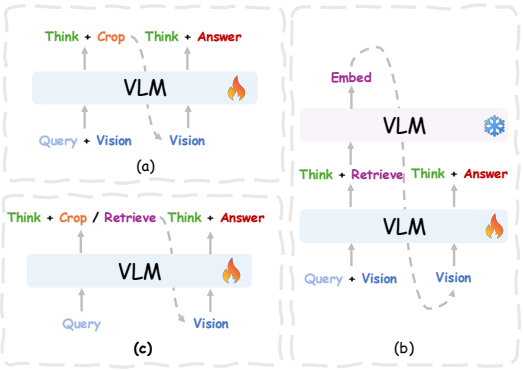



Figure 1: **Think with video paradigms for long-form video understanding.** (a) End-to-end tool-calling MLLMs rely solely on generation-based strategy, iteratively cropping relevant segments. (b) Retrieval-based methods use a frozen model for embedding, decoupling evidence acquisition from optimization. (c) Route2Learn supports both strategies within a single model and learns query-adaptive routing through policy optimization.

A natural approach to optimizing this routing policy is to apply vanilla Group Relative Policy Optimization (GRPO) (Guo et al., 2025) on top of a cold-started model. However, vanilla GRPO faces two challenges. First, the model may exhibit an inherent tool preference bias, which prevents it from exploring both routes. Simply increasing the group size neither guarantees coverage of both strategies nor avoids additional training cost. Second, a single sequence-level advantage is assigned to all generated tokens, failing to distinguish the contribution of tool selection from answer generation.

To address these challenges, Route2Learn is built on a simple yet important insight: ***learning** to route requires comparing both **routes***. By directly evaluating both routes for each query, we can overcome tool preference bias and derive a routing signal from their relative outcomes. To realize this idea, we introduce a two-stage optimization approach. For Supervised Fine-Tuning (SFT), we actively execute both tools for each query and progressively collect trajectories with rejection sampling, providing an implicit signal about which strategies succeed for different queries. We then propose **Counterfactual Routing Policy Optimization (CRPO)** to explicitly optimize routing preferences between the two strategies. For each factual rollout, CRPO constructs a counterfactual rollout by forcing the alternative evidence acquisition strategy. The outcome difference between factual and counterfactual rollouts provides a direct signal for routing, which CRPO further uses to compute distinct advantages for tokens associated with tool selection, enabling precise credit assignment for query-adaptive routing.

Our contributions are threefold:

- We introduce Route2Learn, which, to our knowledge, is the first native single-MLLM framework that learns query-adaptive routing between generation-based and retrieval-based strategies for long-form video understanding.
- We propose a two-stage optimization approach based on the key insight that learning to route requires comparing both routes, which first exposes the model to both strategies during SFT and then explicitly learns their relative preferences through CRPO.
- Extensive experiments on long-form video understanding benchmarks demonstrate that Route2Learn significantly outperforms existing models and learns to adaptively select the more effective evidence-acquisition strategy for queries.

# 2 Related Work

## 2.1 Agentic Video Understanding

Agentic video understanding has evolved along two primary lines. The first line develops natively post-trained single-MLLM methods, which perform generation-based evidence acquisition by dynamically localizing and re-inspecting relevant video segments (Yang et al., 2026a;b; He et al., 2026; Zhang et al., 2026e; Xing et al., 2026; Zhang et al., 2026b; Zeng et al., 2026; Ding et al., 2026). These methods benefit from end-to-end optimization but are limited to a single evidence acquisition strategy. The second line uses LLM-based agents to orchestrate expert modules, including external retrieval models (Yuan et al., 2025; Li et al., 2026a; Qiu et al., 2026; Ma et al., 2026; Wang et al., 2026c). Recently, Route2Look demonstrates the complementarity of generation-based and retrieval-based strategies by routing between them according to the query (Wang et al., 2026c). However, it remains a training-free framework that implements routing through instructions, and still relies on external modules. In contrast, Route2Learn unifies both strategies within a single trainable MLLM and learns query-adaptive routing through direct policy optimization.

## 2.2 RL for Video Understanding

Recent work extends GRPO (Guo et al., 2025) to video understanding with task-specific rewards. Video-R1 introduces a temporal reward by contrasting predictions on ordered and shuffled video inputs (Feng et al., 2026). VITAL calibrates rewards according to task and sample difficulty (Zhang et al., 2026b). FrameThinker combines correctness-gated action rewards with cognitive consistency verification to suppress invalid tool trajectories (He et al., 2026). LongVT uses answer and temporal-IoU rewards, followed by reinforcement fine-tuning on high-reward trajectories (Yang et al., 2026a). Video-o3 balances clue localization and exploration redundancy through trajectory-dependent rewards (Zeng et al., 2026). More closely, some works explicitly consider adaptive tool use. AdaTooler-V rewards tool use based on a precomputed Tool Benefit Score (Wang et al., 2026a), ParaVT encourages beneficial tool use through frame-budget gating (Yang et al., 2026b), VideoSearcher introduces a bell-shaped reward to encourage appropriate tool use (Gao et al., 2026), and AdaVDR employs a redundancy-aware reward to discourage unnecessary calls (Zhang et al., 2026d). These methods primarily optimize tool necessity. In contrast, our CRPO focuses on adaptive selection between complementary tools, beyond deciding whether to invoke a tool. CRPO learns to select between complementary tools by comparing factual rollouts with counterfactual rollouts. It further decouples tool and answer advantages to provide targeted credit assignment for query-adaptive routing.

# 3 Method

## 3.1 Overall Agentic Pipeline

Route2Learn formulates video question answering as an iterative evidence-acquisition process. Given a query $Q$, a single MLLM $\pi_\theta$ progressively gathers visual evidence from the video $V$ through adaptive tool use and then synthesizes the accumulated evidence into a final answer. The overall pipeline is illustrated in Figure 2.

**Evidence-acquisition tools.** Route2Learn equips the model with two complementary tools

$$\mathcal{T} = \{\texttt{crop_video}, \texttt{retrieve}\}.$$

The generation-based tool, `crop_video`, allows the model to inspect one or more temporal intervals:

$$a_{\text{gen}} = \texttt{crop_video}\,(s, e)\,, \tag{1}$$

where $s$ and $e$ denote the start and end timestamps. The tool samples frames from the specified intervals and returns them as visual observations. By adjusting the interval width, the model can inspect local details or broader temporal context. When the query contains explicit temporal cues, the model calls the tool to inspect the corresponding interval directly. Consequently, `crop_video` is particularly suitable for queries involving explicit temporal references, long-range dependencies, or global event understanding.

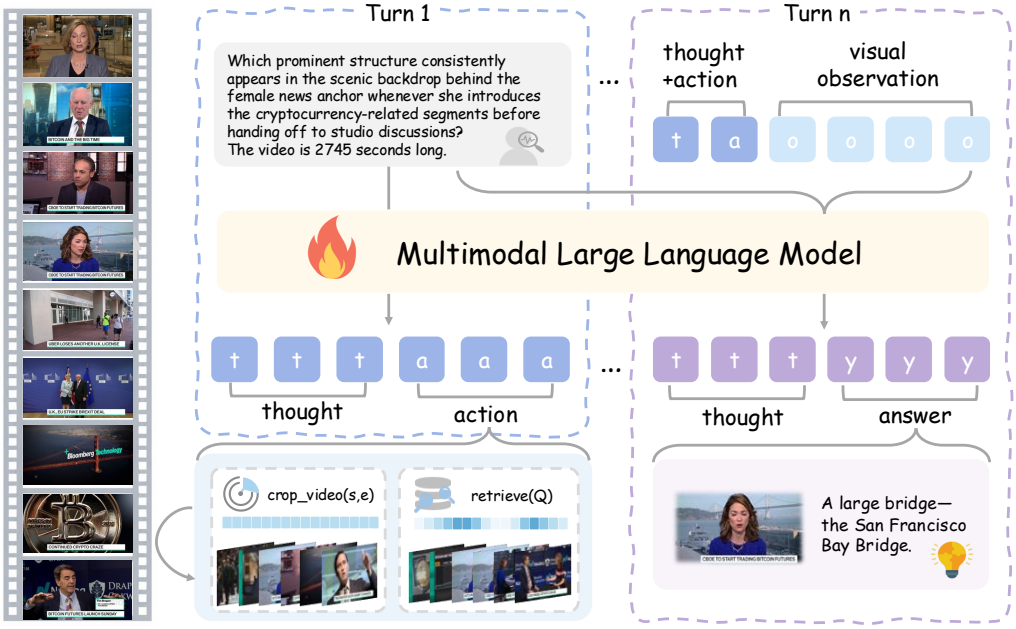



Figure 2: **Overview of Route2Learn.** Given a query and video metadata, a single trainable MLLM iteratively generates thoughts and selects between `crop_video` for temporal inspection and `retrieve` for semantic retrieval. The acquired visual observations are incorporated into the interaction history to support subsequent reasoning and final answer generation.

The retrieval-based tool, `retrieve`, searches the entire video using the query:

$$a_{\text{ret}} = \texttt{retrieve}(Q). \tag{2}$$

Following FreeRet (Zhu et al., 2025), we derive embeddings from the output of the attention sub-layer in the final layer of the MLLM, elicited with a controlled prompt template. The query and candidate frames are encoded by the same video MLLM $\pi_\theta$, and the top-$K$ frames with the highest cross-modal similarities are returned as visual observations. During training, both query and frame embeddings are computed using the current MLLM $\pi_\theta$, allowing the retrieval representations to adapt as the shared model is updated. This tool is well suited to locating localized evidence that is difficult to identify through reasoning-driven temporal navigation alone.

**Interaction protocol.** At each tool-use turn $k$, the model generates a textual thought $t^{(k)}$ followed by a tool-call action $a^{(k)}$, conditioned on the interaction history $H^{(k-1)}$:

$$\left(t^{(k)}, a^{(k)}\right) \sim \pi_\theta\left(\cdot \mid H^{(k-1)}\right), \qquad a^{(k)} \in \{a_{\text{gen}}, a_{\text{ret}}\}. \tag{3}$$

Executing the selected tool on $V$ produces a visual observation $o^{(k)}$, and the resulting thought–action–observation triplet is appended to the history:

$$H^{(k)} = \left(H^{(k-1)}, (t^{(k)}, a^{(k)}, o^{(k)})\right), \qquad H^{(0)} = (Q, V_{\text{meta}}), \tag{4}$$

where $V_{\text{meta}}$ denotes video metadata such as duration. If the model generates a final answer at turn $n$, the interaction terminates with the final thought $t^{(n)}$ and answer $y$, without further tool calls. We denote the sequence of all generated tokens by

$$\tau = \left(t^{(1)}, a^{(1)}, \ldots, t^{(n-1)}, a^{(n-1)}, t^{(n)}, y\right). \tag{5}$$

By integrating adaptive tool routing, iterative evidence acquisition, and answer generation within a single video MLLM, Route2Learn dynamically selects the evidence-acquisition strategy according to each query.

### 3.2 SUPERVISED FINE-TUNING

Direct reinforcement learning can be unstable when the base model has limited experience with native tool calling and multi-turn reasoning (Yang et al., 2026a; Xing et al., 2026). Guided by our insight that *learning to route requires comparing both routes*, the SFT stage exposes the model to successful trajectories from both evidence acquisition strategies.

**Paired Trajectory Construction.** For each query, we use a teacher model to interact with the video through both strategies, producing one generation-based trajectory and one retrieval-based trajectory. Neither the reference answer nor the reference interval is provided during this initial construction. Executing both routes ensures that trajectory collection explores both strategies for the same query, regardless of the teacher's initial tool preference.

**Progressive Construction with Rejection Sampling.** We evaluate each trajectory independently and retain it only if it yields the correct answer. For queries that neither route solves, we progressively strengthen the guidance: first requiring deeper temporal inspection, then providing the reference interval while withholding the reference answer. Queries that remain unsolved are discarded rather than given the correct answer, mitigating hallucinated reasoning and teacher-forcing bias.

Finally, we balance the retained trajectories across the two routes, yielding 11.0K tool-calling trajectories with 5.5K per route. This balance reduces the risk of tool preference bias induced by unequal training exposure. Overall, SFT provides implicit routing supervision and equips the model to execute both strategies. However, because the two routes are supervised independently, SFT does not explicitly optimize the routing preferences, motivating the RL stage.

### 3.3 COUNTERFACTUAL ROUTING POLICY OPTIMIZATION

After SFT, we treat the resulting model as an interactive agent that autonomously selects between `crop_video` and `retrieve` to gather visual evidence before answering the query. We then optimize this tool-using policy through reinforcement learning.

Vanilla GRPO (Guo et al., 2025) samples $G$ rollouts $\{\tau_i\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid V, Q)$. Each rollout is assigned an outcome reward $R_i^{\text{oc}}$ based on answer accuracy $r_{\text{acc}}(\tau_i)$ and format compliance $r_{\text{fmt}}(\tau_i)$ following Yang et al. (2026b). GRPO then computes the group-normalized advantage

$$A_i^{\text{oc}} = \frac{R_i^{\text{oc}} - \text{mean}\big(\{R_j^{\text{oc}}\}_{j=1}^{G}\big)}{\text{std}\big(\{R_j^{\text{oc}}\}_{j=1}^{G}\big) + \epsilon}, \tag{6}$$

which is uniformly assigned to all generated tokens in $\tau_i$. This formulation has two challenges for routing optimization. First, an existing tool preference may prevent the sampled group from covering both strategies. Increasing the group size adds training cost without guaranteeing such coverage. Second, assigning the same trajectory-level advantage to all generated tokens conflates tool selection with answer generation.

To address these challenges, we propose **Counterfactual Routing Policy Optimization (CRPO)**, which explicitly evaluates the alternative route and decouples advantages for tool and answer tokens, as illustrated in Figure 3.

**Counterfactual Rollout Construction.** For each factual rollout $\tau_i$, we construct a counterfactual rollout $\tilde{\tau}_i$ by forcing the alternative tool and regenerating from the same input. The counterfactual rollout is used to compute the tool reward and does not directly contribute to policy gradients.

**Counterfactual Tool Reward.** A factual–counterfactual pair is considered valid when the counterfactual satisfies the interaction constraints, executes the required alternative tool, and produces an extractable answer. Let $\mathcal{P}_{\text{valid}}$ denote the indices of valid pairs within a rollout group. For each valid pair, we define

$$R_i^{\text{tool}} = r_{\text{acc}}(\tau_i) - r_{\text{acc}}(\tilde{\tau}_i), \tag{7}$$

where $r_{\text{acc}}$ evaluates the final answer by exact match against the ground truth. This accuracy difference provides routing supervision without explicit tool-selection labels. We compute $\mu_{\text{tool}}$ and

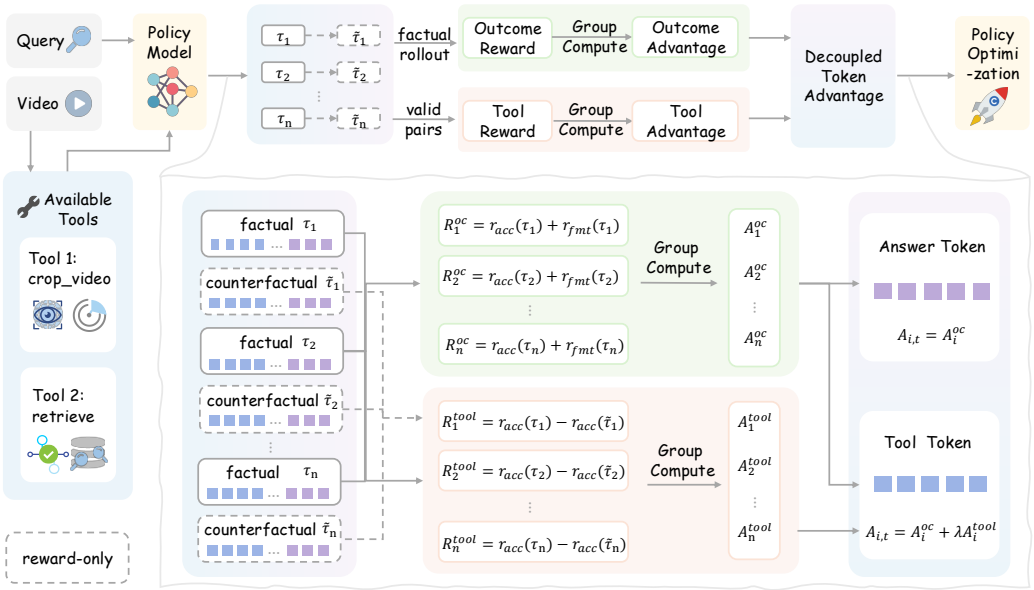



Figure 3: **Overview of CRPO.** Each factual rollout is paired with a counterfactual rollout that forces the alternative tool. Tool advantages are derived from the accuracy gap between each valid factual–counterfactual pair and are applied exclusively to *tool tokens* (blue) alongside the outcome advantage, while *answer tokens* (purple) receive only the outcome advantage. Dashed boxes denote reward-only counterfactual rollouts, which are used to compute tool rewards but do not directly contribute to policy gradients.

$\sigma_{\text{tool}}$ as the mean and standard deviation of tool rewards over all valid pairs, including ties. The tool advantage is

$$A_i^{\text{tool}} = \begin{cases} \dfrac{R_i^{\text{tool}} - \mu_{\text{tool}}}{\sigma_{\text{tool}} + \epsilon}, & i \in \mathcal{P}_{\text{valid}} \text{ and } R_i^{\text{tool}} \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Valid ties participate in normalization but are masked to zero after normalization, while invalid pairs are excluded from normalization and also receive zero tool advantage.

**Decoupled Token Advantage.** To provide targeted credit assignment between tool use and answer generation, CRPO partitions the tokens in each factual rollout into two segments:

$$\tau_i^{\text{tool}} = \left(t_i^{(1)}, a_i^{(1)}, \ldots, t_i^{(n-1)}, a_i^{(n-1)}\right), \quad \tau_i^{\text{ans}} = \left(t_i^{(n)}, y_i\right), \tag{9}$$

where *tool tokens* $\tau_i^{\text{tool}}$ consist of all thought and action tokens generated during tool-use turns in the factual rollout, whereas *answer tokens* $\tau_i^{\text{ans}}$ comprise the final thought and answer. The tool advantage is applied only to the tool tokens:

$$A_{i,t} = \begin{cases} A_i^{\text{oc}} + \lambda A_i^{\text{tool}}, & t \in \tau_i^{\text{tool}}, \\ A_i^{\text{oc}}, & t \in \tau_i^{\text{ans}}, \end{cases} \tag{10}$$

where $t$ indexes response tokens and $\lambda$ balances the two advantages. We optimize the policy on factual rollouts using the following objective:

$$\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|}\min\Big(\rho_{i,t}(\theta)A_{i,t}, \text{clip}(\rho_{i,t}(\theta), 1-\delta, 1+\delta)\, A_{i,t}\Big)\right] - \beta\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{11}$$

where $\rho_{i,t}(\theta)$ is the corresponding importance sampling ratio $\pi_\theta / \pi_{\theta_{\text{old}}}$.

Table 1: **Comparison with existing Video-MLLMs.** The best result is in **bold**. $*$ marks cells withheld due to benchmark–training-data overlap. Unavailable results are denoted by "–". "Avg." is the macro-average over the four benchmarks.

| Method | LVBench | VideoMME | MLVU | LongVideo-Bench | Avg. |
|---|---|---|---|---|---|
| | | w/ sub | M-Avg | | |
| *Proprietary Video-MLLMs* | | | | | |
| GPT-4o (Hurst et al., 2024) | 34.7 | 77.2 | 64.6 | 66.7 | 60.8 |
| Gemini 1.5 Pro (Team et al., 2024) | 33.1 | 81.3 | 74.3 | 64.4 | 63.3 |
| *Open Instruct Video-MLLMs* | | | | | |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 32.2 | 64.5 | 47.8 | 46.4 | 47.7 |
| *Open Reasoning Video-MLLMs* | | | | | |
| VideoChat-R1-7B (Li et al., 2025) | 23.8 | 58.2 | 58.7 | 49.2 | 47.5 |
| Video-Thinker-7B (Wang et al., 2025a) | $*$ | 65.3 | 65.2 | 56.0 | – |
| Video-R1-7B (Feng et al., 2026) | 36.9 | 66.0 | 61.6 | 57.4 | 55.5 |
| VideoRFT-7B (Wang et al., 2026b) | 38.0 | 65.6 | 44.9 | 55.1 | 50.9 |
| Time-R1-7B (Wang et al., 2026d) | 38.2 | 66.2 | 60.5 | 56.0 | 55.2 |
| ReWatch-R1-7B (Zhang et al., 2026a) | 38.5 | 65.0 | 60.1 | 53.6 | 54.3 |
| *Open Agentic Video-MLLMs* | | | | | |
| Qwen3-VL-8B (Bai et al., 2025a) | 33.1 | 68.4 | 58.3 | 52.2 | 53.0 |
| VideoZoomer-7B (Ding et al., 2026) | 22.9 | 48.3 | 46.2 | 39.6 | 39.3 |
| SAGE-7B (Jain et al., 2026) | 31.8 | 52.4 | 49.7 | 37.4 | 42.8 |
| Conan-7B (Ouyang et al., 2026) | 38.2 | 62.8 | 59.2 | 54.5 | 53.7 |
| LongVT-RFT-7B (Yang et al., 2026a) | 37.9 | 66.0 | 59.4 | 54.7 | 54.5 |
| ParaVT-8B (Yang et al., 2026b) | 39.8 | 69.4 | 65.0 | 60.4 | 58.7 |
| **Route2Learn-8B (Ours)** | **50.5** | **70.8** | **70.8** | **61.5** | **63.4** |

# 4 Experiments

## 4.1 Experimental Setup

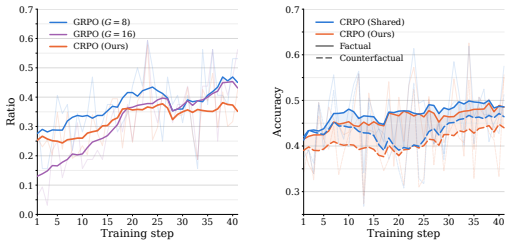
**Implementation Details.** We build Route2Learn on Qwen3-VL-8B-Instruct (Bai et al., 2025a). In the SFT stage, we use Qwen3-VL-235B-A22B (Bai et al., 2025a) as the teacher model, yielding 11.0K tool-calling trajectories on the training queries from Yang et al. (2026a) with an equal balance between the two routes. Following common practice (Yang et al., 2026a;b), we additionally mix in 55.2K samples from Video-R1 (Feng et al., 2026) and LongVideo-Reason (Chen et al., 2026) without tool-calls to preserve general reasoning capability. In the CRPO stage, we train on 2K queries sampled from CG-Bench (Chen et al., 2025b), pre-filtered to remove queries for which all responses sampled from the cold-start policy are incorrect. We set the factual group size $G = 8$, the balance coefficient $\lambda = 1.0$, and the maximum number of tool-calling turns to 3. The model is trained on 8 NVIDIA H200 GPUs. More details are provided in Appendix A.

**Benchmarks and Metrics.** We evaluate on four long-form video understanding benchmarks following the sparse frame sampling protocol used by Yang et al. (2026a;b), reporting MCQ accuracy on VideoMME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2025b), and MLVU (Zhou et al., 2025). These benchmarks cover diverse video domains, durations, and query types, and collectively require both localized evidence discovery and long-range temporal reasoning.

**Baselines.** We compare Route2Learn with three groups of methods: (1) direct-answer models based on instruct backbones with no native thinking pattern, (2) reasoning-enhanced models trained with a chain-of-thought schema, and (3) tool-augmented agentic models with native tool-calling capabilities. GPT-4o and Gemini 1.5 Pro are reported as proprietary reference rows using their officially reported results. We restrict our evaluation to natively post-trained single-MLLM methods, excluding agent frameworks that use external models (Yuan et al., 2025; Chen et al., 2025a; Zhang et al., 2026c; Li et al., 2026a; Wang et al., 2026c) to keep the comparison fair.

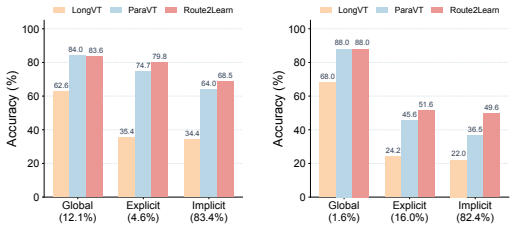
Table 2: **Ablation studies.** Default settings are marked in gray. "Avg." denotes the macro-average over the four benchmarks.

| Setting | LVBench | VideoMME | MLVU | LongVideoBench | Avg. |
|---|---|---|---|---|---|
| *(a) Routing strategy* | | | | | |
| Crop-only | 45.5 | 69.3 | 65.6 | 59.6 | 60.0 |
| Retrieve-only | 47.2 | 70.2 | 68.6 | 56.7 | 60.7 |
| Random routing | 44.6 | 69.9 | 66.4 | 57.2 | 59.5 |
| Adaptive routing | 50.5 | 70.8 | 70.8 | 61.5 | 63.4 |
| *(b) Training stage* | | | | | |
| Qwen3-VL-8B | 33.1 | 68.4 | 58.3 | 52.2 | 53.0 |
| + SFT | 43.7 | 68.0 | 63.3 | 55.9 | 57.7 |
| + SFT + GRPO ($G = 8$) | 47.1 | 69.7 | 68.5 | 60.5 | 61.5 |
| + SFT + GRPO ($G = 16$) | 48.6 | 71.0 | 68.6 | 60.6 | 62.2 |
| + SFT + CRPO ($G = 8$) | 50.5 | 70.8 | 70.8 | 61.5 | 63.4 |
| *(c) Advantage assignment* | | | | | |
| + SFT + CRPO (shared) | 46.9 | 70.2 | 68.0 | 59.6 | 61.2 |
| + SFT + CRPO (decoupled) | 50.5 | 70.8 | 70.8 | 61.5 | 63.4 |




(a) Zero-advantage ratio (b) F vs. CF accuracy

Figure 4: **Training dynamics.** The left and right panels show changes in the zero-advantage ratio and factual (F)-counterfactual (CF) accuracy during training respectively.




(a) MLVU (b) LVBench

Figure 5: **Performance on different query types.** Percentages in parentheses denote the proportion of each type within the corresponding benchmark.

### 4.2 Main Results

Table 1 shows that Route2Learn achieves the highest accuracy among the evaluated open models across all four benchmarks, with an average of 63.4%. It surpasses its Qwen3-VL-8B backbone by +10.4% on average and outperforms ParaVT-8B, the strongest open baseline, by +4.7%. The gains over ParaVT-8B are particularly pronounced on LVBench and MLVU, reaching +10.7% and +5.8%, respectively. Route2Learn also achieves an average accuracy above the proprietary results of GPT-4o and Gemini 1.5 Pro. Overall, these results demonstrate the strong performance of Route2Learn across different benchmarks.

### 4.3 Ablation Study

**Routing strategy.** Table 2(a) shows that retrieve-only performs better on LVBench, VideoMME, and MLVU, whereas crop-only is stronger on LongVideoBench, highlighting their complementary strengths. Adaptive routing outperforms both fixed strategies and random routing, achieving an average accuracy of 63.4% compared with 60.7% for the strongest fixed strategy.

**Training stage.** Table 2(b) shows that reinforcement learning improves the SFT model across all benchmarks, with CRPO outperforming GRPO at $G = 8$. Compared with GRPO at $G = 16$, CRPO achieves stronger or comparable results with only half as many factual rollouts directly contributing policy gradients, reaching an average accuracy of 63.4% over 62.2%. Figure 4(a) further shows that

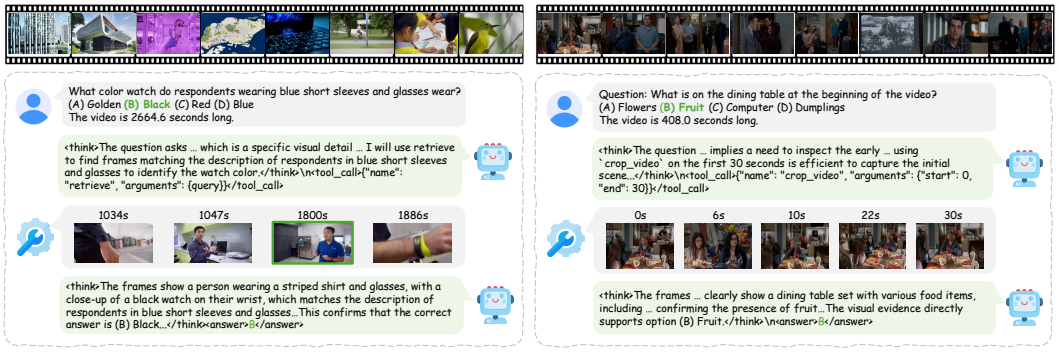



Figure 6: **Case study of query-adaptive routing.** Left: Route2Learn invokes `retrieve` for semantically localized queries, identifying a watch color in a long video. Right: It invokes `crop_video` for temporally anchored queries, inspecting the beginning segment to identify fruit.

CRPO maintains a lower zero-advantage ratio than both GRPO variants in later training, suggesting that counterfactual evaluation provides additional learning signals when outcome advantages are zero.

**Advantage assignment.** Table 2(c) shows that decoupled advantages consistently outperform the shared-advantage variant across all benchmarks, improving average accuracy from 61.2% to 63.4%. As shown in Figure 4(b), both variants improve factual accuracy during training, while the decoupled variant generally maintains a larger factual–counterfactual accuracy gap. Together, these results support decoupled credit assignment for learning more discriminative routing preferences.

### 4.4 Performance on Different Query Types

Following the taxonomy of Wang et al. (2026c), we analyze performance across three query types in Figure 5: global understanding, explicit temporal, and implicit temporal. For a fair comparison, we reproduce two strong baselines, LongVT and ParaVT, under the same settings. Route2Learn achieves the highest accuracy on both explicit and implicit temporal queries across MLVU and LVBench, while remaining competitive on global understanding. In particular, Route2Learn improves over the strongest baseline on the dominant and challenging implicit-temporal queries by +4.5% and +13.1%, respectively. While both baselines begin with a video overview, Route2Learn selects targeted temporal inspection or semantic retrieval according to the query.

### 4.5 Case Study

Figure 6 shows representative cases of query-adaptive routing. In the left example, the query asks about a localized attribute in an extremely long video. Route2Learn invokes `retrieve` to search for semantically relevant frames across the entire video and successfully identifies the watch as black. In the right example, the query explicitly provides a clear temporal anchor. Route2Learn therefore invokes `crop_video` to inspect the first 30 seconds directly and correctly identifies fruit on the dining table. More cases are provided in Appendix E.

## 5 Conclusion

In this work, we propose Route2Learn, an end-to-end native agentic framework that learns query-adaptive routing between generation-based and retrieval-based strategies within a single video MLLM for long-form video understanding. Guided by the insight that learning to route requires comparing both routes, we introduce a two-stage optimization approach with SFT for paired trajectory exposure and CRPO for explicitly optimizing routing preferences. Experiments on four long-form video understanding benchmarks demonstrate consistent improvements over evaluated models, highlighting the promise of query-adaptive routing for agentic long-form video understanding.

### AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic training data. Specifically, we used Qwen3-VL-235B-A22B to generate tool-calling trajectories for supervised fine-tuning, as described in Sections 3.2 and 4.1 and Appendix C. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, translating text, cleaning or reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Providing critical ingredients for mathematical proofs and assisting with proof writing are not applicable to this work. Additionally, we used generative AI tools for manuscript writing and polishing, including sentence rephrasing, grammar checking, and improvements to readability, clarity, and flow. We have reviewed all AI-assisted work and revised the generated or polished text where necessary to ensure accuracy and consistency with the scientific content of the paper. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

Our research is conducted in accordance with established ethical practices, with particular attention to the responsible use of data. The datasets used in this study are established academic datasets and benchmarks that are commonly used in the video understanding and multimodal learning communities. We do not collect new private or sensitive personal information, and our experiments do not involve human-subject studies. We follow the usage conditions and licenses of the datasets employed in this work.

### REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of the research presented in this paper. Section 4.1 reports the main experimental settings, model initialization, training data, and evaluation protocol. Appendix A provides the tool configurations and training hyperparameters. Appendix C describes the evaluation benchmarks and the construction of the SFT and RL data. Appendix F provides the prompts used for SFT, RL, evaluation, and retrieval. Furthermore, we will publicly release the training data, code, and model weights to facilitate reproduction and further research upon publication.

## REFERENCES

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*, 2025a.

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

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

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025b.

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

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

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

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.

Zhenkun Gao, Yicheng Bao, Jinlong Peng, Xueheng Li, Theo Huang, Bangwei Liu, Kunquan Li, Zhenye Gan, Tao Hu, Chengjun Xie, et al. Videosearcher: Empowering video deep research with multi-tool agentic reasoning via reinforcement learning. *arXiv preprint arXiv:2607.02927*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

Chenglin Li, Qianglong Chen, Feng Han, Yikun Wang, Xingxi Yin, Yan Gong, Ruilin Li, Yin Zhang, and Jiaqi Wang. Videothinker: Building agentic videollms with llm-guided tool reasoning. *arXiv preprint arXiv:2601.15724*, 2026a.

Miaoyu Li, Qin Chao, and Boyang Li. Two causally related needles in a video haystack. *Advances in Neural Information Processing Systems*, 38, 2026b.

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

Martin Q Ma, Yuxiao Qu, Aditya Agrawal, Willis Guo, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Act2see: Emergent active visual perception for video reasoning. *arXiv preprint arXiv:2605.01657*, 2026.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41089–41099, 2026.

Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. Videoseal: Mitigating evidence misalignment in agentic long video understanding by decoupling answer authority. *arXiv preprint arXiv:2605.12571*, 2026.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

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

Chaoyang Wang, Kaituo Feng, Dongyang Chen, Zhongyu Wang, Zhixun Li, Sicheng Gao, Meng Meng, Xu Zhou, Manyuan Zhang, Yuzhang Shang, et al. Adatooler-v: Adaptive tool-use for images and videos. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18058–18072, 2026a.

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

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

Tianyue Wang, Xuying Wu, Yuxiang Ma, Ruiming Liang, Jiaxuan Kang, Yanchao Hao, Zheng Wei, Leigang Qu, Haiyun Guo, and Jinqiao Wang. Routing before looking: Query-adaptive evidence acquisition for long-form video understanding. *arXiv preprint arXiv:2608.20805*, 2026c.

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

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.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026d.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9088–9099, 2026e.

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

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *Advances in Neural Information Processing Systems*, 38:143297–143330, 2026.

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

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026a.

Zuhao Yang, Kaichen Zhang, Sudong Wang, Keming Wu, Zhongyu Yang, Bo Li, Xiaojuan Qi, Shijian Lu, Xingxuan Li, and Lidong Bing. Paravt: Taming the tool prior paradox for parallel tool use in agentic video reinforcement learning. *arXiv preprint arXiv:2605.20342*, 2026b.

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.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. Videoexplorer: Think with videos for agentic long-video understanding. *arXiv preprint arXiv:2506.10821*, 2025.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026.

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

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026b.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026c.

Xintong Zhang, Xiaomeng Fan, Shilin Yan, Ekko He, Zicheng Liu, Zijian Zou, Guannan Zhang, Yuwei Wu, Zhi Gao, and Hongwei Xue. Adavdr: Adaptive tool use and reflection for video deep research. *arXiv preprint arXiv:2608.25559*, 2026d.

Yaolun Zhang, Ruohui Wang, Jiahao Wang, Yepeng Tang, Xuanyu Zheng, Haonan Duan, Hao Lu, Hanming Deng, and Lewei Lu. Eva: Efficient reinforcement learning for end-to-end video agent. *arXiv preprint arXiv:2603.22918*, 2026e.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

Yuhan Zhu, Xiangyu Zeng, Chenting Wang, Xinhao Li, Chunxu Liu, Yicheng Xu, Ziang Yan, Yi Wang, and Limin Wang. Freeret: Mllms as training-free retrievers. *arXiv preprint arXiv:2509.24621*, 2025.

# APPENDIX

## APPENDIX CONTENTS

**A More Implementation Details** **14**
A.1 Tool configuration . . . . . 14
A.2 Training hyperparameters . . . . . 14

**B Sensitivity Analysis** **14**

**C Datasets and Benchmarks** **15**
C.1 Details of Benchmarks . . . . . 15
C.2 Details of SFT Data . . . . . 16
C.3 Details of RL Data . . . . . 16

**D Efficiency Analysis** **17**

**E Additional Case Studies** **17**
E.1 Successful cases . . . . . 17
E.2 Failure case . . . . . 17

**F Prompts** **17**

## A MORE IMPLEMENTATION DETAILS

### A.1 TOOL CONFIGURATION

A `crop_video` call returns 64 frames sampled from the predicted temporal interval. A `retrieve` call searches a candidate pool constructed by sampling the video at 1.0 FPS, subject to a maximum of 768 candidate frames, and returns the top-$K$ frames with $K = 64$. Thus, both tools return 64 frames per call, while differing in how they select the visual evidence. We allow at most three tool-calling turns per query, with up to three `crop_video` calls per turn.

### A.2 TRAINING HYPERPARAMETERS

Tables 3 and 4 summarize the hyperparameters for SFT and RL, respectively. We initialize the cold-start policy from Qwen3-VL-8B-Instruct. Before RL, we filter out queries for which all responses sampled from the cold-start policy are incorrect. This filtering aims to reduce factual–counterfactual pairs in which both rollouts are incorrect and therefore receive zero tool advantage. During CRPO, we sample $G = 8$ factual rollouts and generate eight corresponding counterfactual rollouts, one for each factual rollout. The counterfactual rollouts are used to compute tool rewards and do not directly contribute policy gradients.

## B SENSITIVITY ANALYSIS

Table 5 reports the performance of Route2Learn trained with different values of $\lambda$. These results suggest that Route2Learn is relatively robust to the choice of $\lambda$ within the evaluated range.

Table 3: Key hyperparameters for SFT.

| Name | Value |
|---|---|
| Finetuning type | Full |
| Freeze vision tower | False |
| Freeze multi-modal projector | False |
| Freeze language model | False |
| Cutoff len | 32768 |
| image/video max pixels | 50176 |
| Frames per tool call | 64 |
| Batch size per device | 1 |
| Gradient accumulation steps | 2 |
| Learning rate | 2.0e-5 |
| LR scheduler type | cosine |
| Warmup ratio | 0.05 |
| Epochs | 1 |

Table 4: Key hyperparameters for RL.

| Name | Value |
|---|---|
| Algorithm | CRPO |
| Max tool rounds | 3 |
| Agent loop | Tool agent |
| Rollouts per group | 8 |
| Train batch size | 32 |
| Mini batch size | 8 |
| Micro batch size per GPU | 2 |
| Learning rate | 1.0e-6 |
| KL loss coefficient | 0.001 |
| Entropy coefficient | 0.0 |
| Max prompt length | 2048 |
| Max response length | 20480 |
| Epochs | 1 |

Table 5: Sensitivity analysis of $\lambda$. "Avg." denotes the macro-average over the four benchmarks. All other settings are held fixed.

| $\lambda$ | LVBench | VideoMME | MLVU | LongVideoBench | Avg. |
|---|---|---|---|---|---|
| 0.3 | 50.4 | 68.2 | 71.9 | 61.9 | 63.1 |
| 0.5 | 50.9 | 70.4 | 70.6 | 62.9 | 63.7 |
| 1.0 | 50.5 | 70.8 | 70.8 | 61.5 | 63.4 |

# C Datasets and Benchmarks

## C.1 Details of Benchmarks

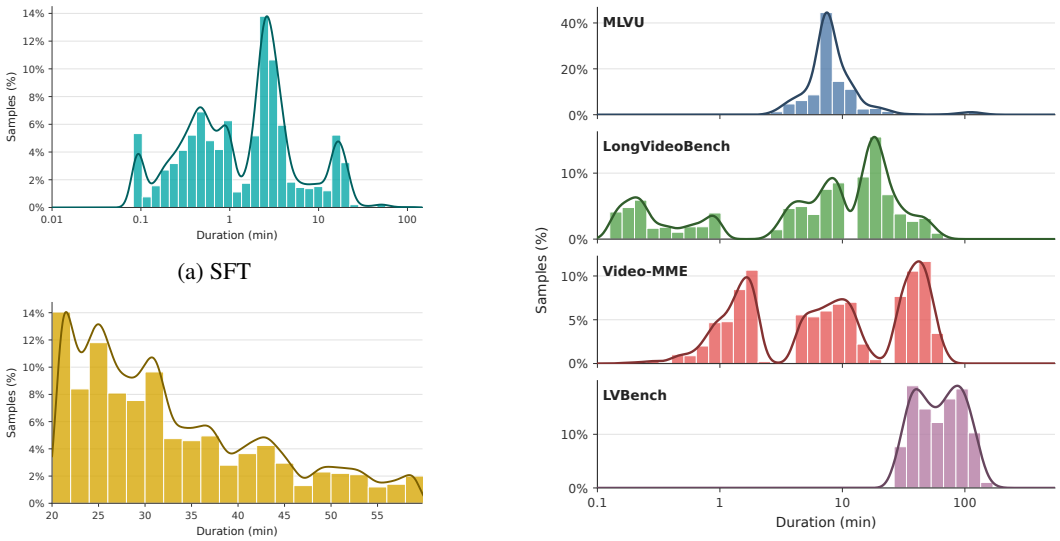
We evaluate Route2Learn on four complementary benchmarks that cover diverse video domains, temporal scales, and reasoning requirements. Figure 7(c) presents their video-duration distributions. Together, these benchmarks assess both the ability to locate fine-grained evidence and the ability to integrate information across extended temporal contexts.

**LVBench.** LVBench (Wang et al., 2025b) focuses on extremely long video understanding, with an average video duration of over one hour. It evaluates six core capabilities: entity recognition, event understanding, key information retrieval, temporal grounding, reasoning, and summarization. Its extended video durations make it particularly useful for evaluating evidence acquisition when relevant events are sparsely distributed within a long video.

**VideoMME.** VideoMME (Fu et al., 2025) contains 900 videos and 2,700 manually annotated question–answer pairs, covering six domains and 30 subcategories. Video durations range from 11 seconds to one hour, with separate short-, medium-, and long-video groups. This diversity enables evaluation across both content domains and temporal scales.

**MLVU.** MLVU (Zhou et al., 2025) evaluates multiple aspects of long video understanding using videos ranging from approximately three minutes to over two hours. It covers diverse genres, including movies, surveillance footage, egocentric videos, and gameplay. Its tasks assess holistic understanding, single-detail understanding, and reasoning over multiple details. We evaluate its multiple-choice tasks and report M-Avg, the mean accuracy across these tasks.

**LongVideoBench.** LongVideoBench (Wu et al., 2024) evaluates long-context understanding using videos with subtitles and durations of up to one hour. It contains 6,678 human-annotated multiple-choice questions across 17 fine-grained categories. The benchmark emphasizes referring reasoning, in which a model must identify the context referenced by a question and reason about the relevant




(a) SFT

(b) RL

(c) Evaluation benchmarks

Figure 7: Video-duration distributions for the SFT data, RL data, and evaluation benchmarks.

visual or textual details. It therefore tests both evidence localization and reasoning over temporally distributed information.

### C.2 Details of SFT Data

Table 6 summarizes the progressive construction of the tool-calling SFT trajectories. The initial construction produces 4,744 successful generation-based trajectories and 5,482 successful retrieval-based trajectories. For queries solved by neither route, using a narrower temporal window produces 72 additional trajectories, while supplying the reference interval without revealing the answer produces another 805 trajectories. This results in 5,621 generation-based trajectories before balancing. We then remove 139 generation-based trajectories through downsampling, yielding a balanced set of 10,964 trajectories with 5,482 trajectories for each route.

Table 6: Construction of the tool-calling SFT trajectories. Negative values indicate trajectories removed during balancing.

| Construction stage | Gen.-based | Ret.-based | Total |
|---|---|---|---|
| Initial construction | 4,744 | 5,482 | 10,226 |
| Narrower temporal window | 72 | – | 72 |
| Reference interval | 805 | – | 805 |
| Balance downsampling | −139 | – | −139 |
| **Final** | **5,482** | **5,482** | **10,964** |

As shown in Figure 7(a), the video-related SFT samples span durations from 4.26 seconds to 145.89 minutes, covering short clips, minute-scale videos, and videos exceeding two hours. This broad temporal coverage exposes the model to evidence acquisition and reasoning at different temporal scales, providing a foundation for subsequent routing optimization.

### C.3 Details of RL Data

We deliberately select RL queries associated with videos of approximately 20–60 minutes, as illustrated in Figure 7(b). Compared with the predominantly shorter SFT samples, these videos provide substantially longer temporal contexts in which to optimize routing decisions. This selection is intended to encourage the policy to locate relevant evidence within extended videos, supporting gener-

alization to hour-scale inputs. It provides a practical intermediate training range between short-video supervision and evaluation on videos lasting an hour or longer.

# D Efficiency Analysis

Table 7 reports the average accuracy and online inference latency across the four benchmarks. Route2Learn achieves the highest average accuracy of 63.4% and the lowest online latency of 7.8 seconds per query. For each video, frame embeddings are computed once to construct the retrieval pool and reused across subsequent queries. This offline preprocessing takes 8.9 seconds per video on average and is not included in the online inference latency.

Table 7: Accuracy and online inference latency of different models. Results are averaged over the four benchmarks.

| Model | Avg. Acc. (%) ↑ | Latency (s/query) ↓ |
|---|---|---|
| LongVT | 54.5 | 22.1 |
| ParaVT | 58.7 | 13.0 |
| Route2Learn (Ours) | **63.4** | **7.8** |

# E Additional Case Studies

## E.1 Successful cases

Figures 8–10 illustrate the complementary roles of the two tools. In Figure 8, `retrieve` locates a camouflaged toad at approximately 4,873 seconds in a 6,028.7-second video, enabling accurate identification of a localized cue without a temporal reference. In Figure 9, the explicit semantic target "javelin throw" guides retrieval to three distinct occurrences around 237, 331, and 368 seconds, supporting the correct count. In contrast, answering the question about "the last song" in Figure 10 requires temporal context: the model first obtains a coarse overview with `crop_video`, then inspects 3,505.5–3,562.9 seconds to determine the sky color.

## E.2 Failure case

Figure 11 illustrates a remaining limitation in temporal reasoning. Although retrieval locates abseiling, javelin throw, zumba, and playing trombone at 3,126, 5,735, 5,863, and 6,410 seconds, respectively, the model selects an ordering inconsistent with these timestamps.

# F Prompts

We use the same system prompt throughout SFT, RL, and evaluation, as shown in Figure 12. For counterfactual rollout generation, we append the following constraint to the system prompt:

> For this question, you MUST use the {`tool`} tool for evidence gathering. Do NOT call {`other`} under any circumstances. Every other instruction above remains in effect.

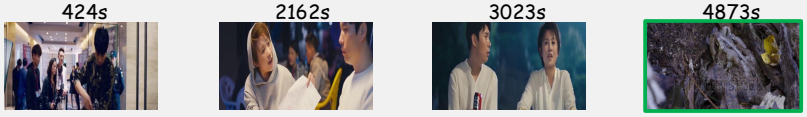
Here, {`other`} denotes the tool used in the first tool call of the factual rollout, and {`tool`} denotes the alternative tool. The frame-side and query-side prompts used by `retrieve` are provided in Figure 13.



What is the American toad's strategy to avoid detection in the video?
(A) Camouflage (B) Fighting (C) Hiding (D) Running away
The video is 6028.7 seconds long.

<think>The question asks about the American toad's strategy to avoid detection, which implies observing its behavior in the video. Since the video is extremely long (6028.7 seconds) and the behavior could occur anywhere, using `retrieve` is optimal to quickly locate the most relevant frames showing the toad's avoidance strategy. I will use `retrieve` to find frames that best illustrate how the toad avoids detection.</think>\n<tool_call>{"name": "retrieve", "arguments": {query}}</tool_call>

424s 2162s 3023s 4873s



<think>The frames show an American toad camouflaged among leaves and soil, blending into its surroundings, which directly supports the idea of camouflage as its strategy to avoid detection. This visual evidence contradicts options B (Fighting), C (Hiding), and D (Running away), as no such behaviors are depicted. My earlier reasoning to use `retrieve` was correct, as it efficiently located the relevant frame showing the toad's camouflage. The evidence is sufficient to confidently select option A.</think>\n<answer>A</answer>

Figure 8: **Localized evidence retrieval.** Route2Learn retrieves a frame showing the toad blending into its surroundings and correctly identifies camouflage as its strategy for avoiding detection.

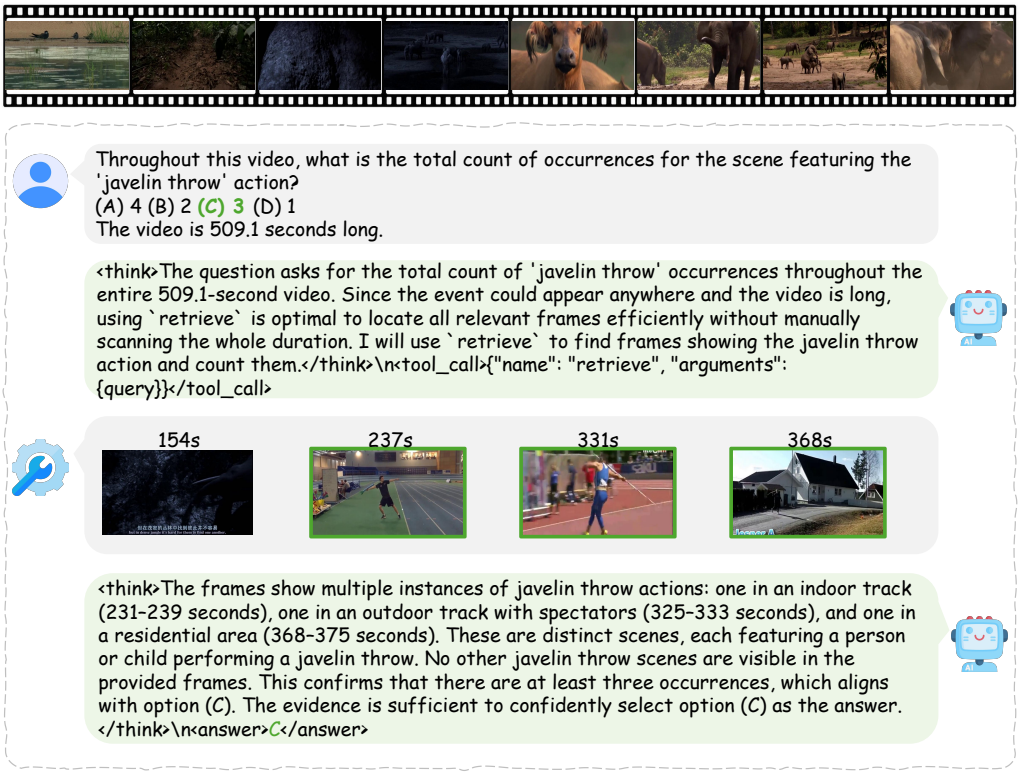



Figure 9: **Semantic retrieval for action counting.** Route2Learn retrieves three distinct occurrences of the specified action, "javelin throw," and correctly determines their count.

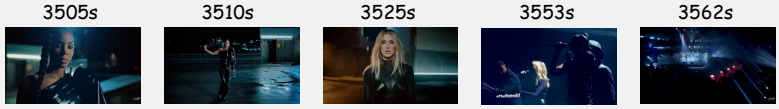


What color is the sky in the last song? (A) Black (B) White (C) Blue (D) Red
The video is 3620.4 seconds long.

<think>I will first watch the entire video from start to finish to identify when the last song occurs and observe the sky's color during that segment.</think>\n<tool_call>{"name": "crop_video", "arguments": {"start": 0.0, "end": 3620.4}}</tool_call>




<think>The video is a compilation of performances from "The Voice" blind auditions, featuring multiple international versions (UK, Belgium, France, etc.) with diverse artists, judges, and stage designs. The scenes alternate between close-ups of singers, wide shots of the stage with audience reactions, and shots of judges in their iconic red chairs. The lighting varies dramatically across segments — from warm reds and purples to cool blues and stark whites — reflecting the different production styles and emotional tones of each performance. The final segment, around 3505.5 seconds, transitions to a dark, atmospheric outdoor scene with a lone figure walking down a narrow alley under a dimly lit sky, which appears to be the last song or performance in the compilation. This final scene's sky color is not clearly visible due to the low light and focus on the figure, but it appears to be a deep, dark blue or black, suggesting nighttime. To confirm the exact color of the sky in this final moment, I will zoom into the [3505.5, 3562.9] segment for a closer look.</think>\n<tool_call>{"name": "crop_video", "arguments": {"start": 3505.5, "end": 3562.9}}</tool_call>




<think>The frames from 3505.5 to 3562.9 show a dark, atmospheric outdoor scene with a lone figure walking down a narrow alley under a dimly lit sky, which appears to be the final performance in the video. The sky is not clearly visible due to the low light and focus on the figure, but the overall lighting suggests a deep, dark blue or black, consistent with nighttime. This confirms my earlier reasoning that the final segment is set outdoors at night, and the sky color is not a bright blue or red as seen in earlier studio scenes. The evidence is sufficient to conclude that the sky in the last song is black, as the darkness and lack of visible sky color point to a night setting.</think>\n<answer>A</answer>

Figure 10: **Progressive temporal inspection.** Route2Learn first obtains a global overview, then inspects a narrower interval near the video's end to answer the question about the last song.

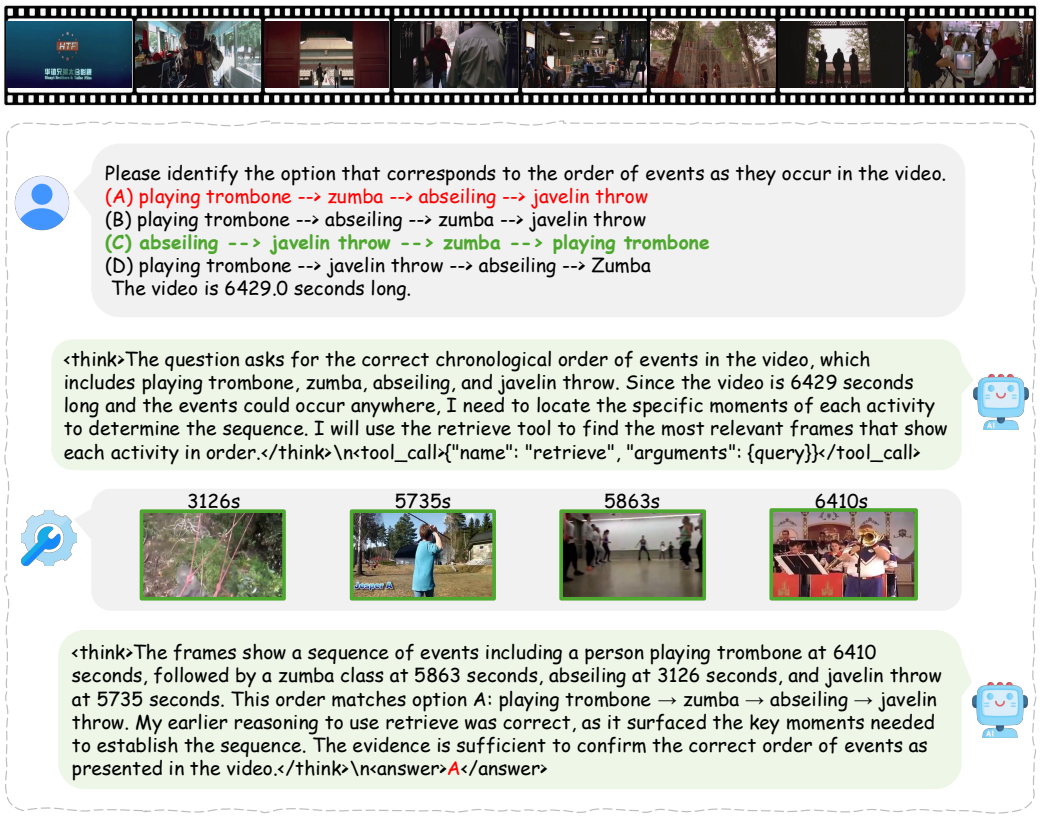



Figure 11: **Failure in temporal reasoning.** Despite retrieving all four queried events, Route2Learn fails to order them according to their timestamps and selects an incorrect answer.

**Prompt for SFT/RL/Eval**

You are a video understanding agent that answers questions about long videos. You are given only the query (no frames); gather visual evidence with tools.

# Workflow
1. **Think**: Analyze the question and decide what evidence you need and which tool fits.
2. **Call tools**:
  - `crop_video`: returns frames from a time range: a narrow window for a specific moment or fine detail, or a wide range (e.g. crop_video(0, <duration>)) for overall content or temporal order. If the question names a time or range (e.g. "between 210 and 260 seconds"), crop exactly that window.
  - `retrieve`: searches the whole video using the question and returns the most relevant frames. Use it to locate a specific detail, object, or moment that could be anywhere (needle-in-haystack).
3. **Answer**: Synthesize your observations into a concise answer.

# Format
Always structure your response as:
<think>your reasoning here</think>
<tool_call>{"name": "crop_video", "arguments": {"start": ..., "end": ...}}</tool_call>
<tool_call>{"name": "retrieve", "arguments": {}}</tool_call>
(use ONLY one tool type per turn; repeat crop_video if needed for multiple segments)
<answer>your final answer</answer>

# Important
- Use at least one tool before answering. Keep answers concise and factual.
- ONLY use `<tool_call>` with JSON format for tool calls. Do NOT use `<tool_code>`, Python function syntax, or any other format.

Figure 12: **System prompt for SFT, RL, and evaluation.** The prompt defines the tool-use workflow, provides guidance for selecting between temporal cropping and semantic retrieval, and specifies the response format.

**Prompt for retrieve (frame side)**

You are required to determine whether the video frame matches the visual content needed to answer the question. <video frame>: [IMG]
Based on the video frame, output exactly one word following these rules:
- The word must capture the core semantics of the frame.
- Do not use function words, symbols, or incomplete fragments.

**Prompt for retrieve (query side)**

You are required to determine whether the video frame matches the visual content needed to answer the question. <query about a video>: [Query]
Based on the question, output exactly one word following these rules:
- The word must capture the core visual semantics needed to answer the question.
- Do not use function words, symbols, or incomplete fragments.

Figure 13: **Prompts for retrieval.** Frame-side and query-side prompts elicit the core visual semantics of candidate frames and queries, respectively, to support frame–query matching.