

Video-MTR: Reinforced Multi-Turn Reasoning for Long Video Understanding

Yuan Xie¹ Tianshui Chen^{2,3} Zheng Ge⁴ Lionel Ni¹

Abstract

Long-form video understanding remains a formidable challenge due to the complexity of modeling long-range temporal dependencies and multi-event narratives. Existing methods often rely on static reasoning or external Visual-Language Models (VLMs), resulting in high computational complexity and sub-optimal performance. In this paper, we propose Video-MTR, a reinforced multi-turn reasoning framework that operates solely through data-efficient, pure RL post-training. Video-MTR reformulates video understanding as a dynamic decision-making process, where the agent iteratively selects key segments conditioned on the evolving context of previously processed frames and the query. To ensure effective intermediate reasoning and training stability, we introduce a novel gated bi-level reward system, which synergizes trajectory-level rewards (answer correctness) with turn-level rewards (frame-query relevance). This mechanism eliminates the need for data-intensive supervised fine-tuning, thereby substantially reducing reliance on large-scale datasets. Remarkably, Video-MTR achieves competitive or superior performance using only $\sim 8K$ training samples, compared to existing approaches that require 257K to 4.4M examples. Extensive experiments on benchmarks including VideoMME, MLVU, LongVideoBench, LVBench, and EgoSchema demonstrate that Video-MTR surpasses state-of-the-art methods in both accuracy and efficiency. Code is available at <https://github.com/Xyuan13/Video-MTR>.

¹The Hong Kong University of Science and Technology (Guangzhou) ²Guangdong University of Technology ³X-Era AI Lab ⁴Stepfun. Correspondence to: Tianshui Chen <tianshuichen@gmail.com>.

1. Introduction

As a foundational computer vision task, video understanding finds widespread applications in numerous domains ranging from intelligent surveillance, content-based retrieval, to autonomous driving. With the explosive growth of user-generated videos and the ubiquity of cameras in daily life, the demand for robust and scalable video-understanding tools has grown substantially. Owing to the advanced reasoning capabilities, Multimodal Large Language Models (MLLMs) (Dai et al., 2023; Wu & Xie, 2024; Weng et al., 2024; Chen et al., 2024b) have demonstrated breakthroughs in visual understanding tasks for images and short videos in recent years. However, long-form video understanding (LVU), characterized by multiple events and long-range temporal dependencies, still presents significant challenges.

Existing approaches (Wang et al., 2024c; Lin et al., 2023; Feng et al., 2025) either employ instruction tuning or recently, integrate reinforcement learning to adapt current MLLMs for long-term temporal reasoning. However, these methods primarily transfer training paradigms designed for language and image modalities, relying on a static reasoning approach that generates predictions based on a fixed, uniform set of sampled frames in a single turn. This single-turn, uniform sampling strategy creates a bottleneck for downstream reasoning tasks when dealing with long-form videos, as it risks omitting critical information due to the extended video duration. Alternatively, other approaches (Fan et al., 2024; Wang et al., 2024b; Ma et al., 2025) explore the agentic paradigm, where large language models (LLMs) serve as agents, utilizing external visual-language models (VLMs) (Radford et al., 2021; Zhao et al., 2023) to identify key video segments. These methods depend on pre-trained VLMs and carefully designed pipelines. While they achieve superior performance, they are hindered by high complexity due to the reliance on heterogeneous external components and sub-optimal tool usage strategies, as they lack end-to-end training.

In this work, we propose Video-MTR, a reinforced multi-turn reasoning framework that leverages the intrinsic capabilities of MLLMs, for iterative key video segment selection and question comprehension within a unified model. Video-MTR builds on an open-source visual language backbone, Qwen2.5-VL-7B (Bai et al., 2025), synergized with rein-

forcement learning based on a carefully-designed bi-level reward that provides fine-grained supervision. Compared to existing paradigms, Video-MTR offers two distinct benefits: 1) It eliminates reliance on external VLMs and carefully designed pipelines; 2) It bypasses the data-intensive supervised fine-tuning process, enabling pure RL training that reduces dependency on large-scale datasets.

Formally, Video-MTR is trained to develop iterative video reasoning capabilities through an end-to-end reinforcement learning strategy. However, current reward systems based solely on answer accuracy offer limited guidance for intermediate video segment selection, particularly in complex long videos. To address this challenge, we introduce a novel gated bi-level reward system, consisting of trajectory-level rewards based on answer correctness and turn-level rewards that capture frame-query relevance. This reward system relies on key segment annotations for turn-level rewards and the final answer for trajectory-level rewards. To enable this, we leverage the limited-scale QA-grounded corpus and augment it with a curated video temporal grounding dataset, using a tailored curation pipeline to align the original annotations with our QA-centric paradigm. Moreover, to maintain video understanding as the primary optimization objective, we anchor frame-level rewards exclusively to final answer correctness, enforcing that intermediate operations must genuinely contribute to the core task. Leveraging such carefully designed rewards that provide fine-grained and task-oriented supervision, Video-MTR eliminates the need for data-intensive supervised fine-tuning process and depend solely on pure, robust, and data-efficient RL training, substantially alleviating reliance on large-scale datasets. Remarkably, Video-MTR achieves competitive or superior performance with only about 8K samples, compared with existing approaches that typically require 257K to 4.4M examples.

The contributions of this work are three-fold. First, we introduce Video-MTR, a reinforced multi-turn reasoning framework that pioneers a data-efficient RL post-training paradigm for long-form video understanding, enabling iterative video segment selection and question comprehension within a unified model. Second, we propose a novel gated bi-level reward mechanism that combines trajectory-level answer correctness with turn-level frame relevance derived from repurposed temporal grounding data. This design provides fine-grained and task-oriented supervision to facilitate effective and stable training. Finally, we conduct extensive experiments on several video understanding benchmarks, including VideoMME (Fu et al., 2025), MLVU (Zhou et al., 2024), LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2024a) and EgoSchema (Mangalam et al., 2023), demonstrating the effectiveness and robustness of Video-MTR. Codes, trained models, and dataset will be released for further research.

2. Related works

2.1. MLLMs for Video Understanding

Building on image MLLMs’ visual reasoning capabilities, researchers develop temporal extensions for video understanding. However, long-form videos remain challenging due to their extended duration exceeding contemporary MLLMs’ context windows. Approaches like Video-LLaVA (Lin et al., 2023), ShareGPT4Video (Chen et al., 2024a), InternVideo2 (Wang et al., 2024c) and Video-R1 (Feng et al., 2025) still resort to uniformly sampling the entire video and rely on post-training with large-scale video-instruction data to boost reasoning abilities. Yet the inevitable loss of information at the input stage creates a performance ceiling. Other approaches explicitly address this bottleneck. One category of methods, exemplified by LongVA (Zhang et al., 2024), LLaMA-VID (Li et al., 2024c), Kangaroo (Liu et al., 2024) and Video-XL (Shu et al., 2025), employs token compression techniques to extend context windows, enabling direct processing of hour-long videos. However, this approach floods the model with redundant information and sacrifices interpretability. Another category, like VideoAgent (Wang et al., 2024b), VideoMemAgent (Fan et al., 2024) and DrVideo (Ma et al., 2025) adopts agent mechanisms (Li et al., 2023; Wu et al., 2023) that dynamically integrate external tools, including video captioning, video object tracking, and key-frame search, through single-turn or multi-turn iterations. Despite outperforming uniform sampling baselines, these systems exhibit high complexity from heterogeneous external components and suboptimal tool utilization due to the absence of end-to-end training.

2.2. MLLMs with Reinforcement Learning

Recent studies (Shen et al., 2025; Meng et al., 2025), inspired by advances in the text domain, have explored reinforcement learning (RL) to improve the reasoning abilities of MLLMs. VLM-R1 (Shen et al., 2025) extends the DeepSeek-R1 paradigm (Guo et al., 2025), showing that an RL-trained MLLM can outperform a supervised fine-tuning baseline and generalize better on visual tasks. DeepEyes (Zheng et al., 2025) incentivizes “thinking with images” over multiple turns via RL. In the video domain, VideoChat-R1 (Li et al., 2025) enhances spatio-temporal perception through reinforcement fine-tuning (RFT) with GRPO, while Video-R1 (Feng et al., 2025) employs a tailored T-GRPO algorithm to emphasize temporal cues. However, these methods primarily target static images or short clips, leaving long-form video understanding largely undressed.

3. Method

In this section, we first introduce the formulation of Video-MTR as a multi-turn interactive reasoning task via reinforcement learning (Sec. 3.1). We then describe the gated bi-level reward mechanism (Sec. 3.2). Finally, we detail the optimization strategy (Sec. 3.3), which leverages pure RL to enable data-efficient learning while ensuring robust reasoning performance.

3.1. Overview

We propose Video-MTR, a framework that reconceptualizes long-form video understanding as a multi-turn interactive reasoning task, closely aligned with the way humans process complex visual information. When presented with a video and a question, humans typically begin by forming a holistic understanding of the overall content, then iteratively attend to specific segments to gather more informative details, and finally integrate the accumulated evidence to derive an answer.

To instantiate this reasoning paradigm, we formulate the task as a reinforcement learning problem. In this formulation, the video functions as a dynamic environment that updates the set of observed frames $\mathcal{F}$ in response to retrieval actions. An MLLM serves as the decision-making agent, interacting with the environment through a learned policy π_θ . As illustrated in Figure 1, the agent operates in a multi-turn manner, and at each step it samples an action $a_k \sim \pi_\theta(\cdot | s_k)$ to either retrieve additional frames or produce the final answer. The state s_k is a multimodal context that concatenates (i) the last w interactions and (ii) the currently observed frames, providing both temporal history and updated visual evidence, and can be represented as

$$s_k = (\mathcal{F}_{k-w}, x_{k-w}, y_{k-w}, \dots, \mathcal{F}_{k-1}, x_{k-1}, y_{k-1}, \mathcal{F}_k, x_k)$$

where x is the text instruction, $\mathcal{F}$ is the set of observed frames, y is the generated response that consists of reasoning process and executable action a . The environment is initialized by uniformly sampling n_0 frames to form $\mathcal{F}_0$ from the whole video. Thereafter, the environment responds to each retrieval action with a new set of frames that become the observation for the next turn. The agent may execute multiple retrieval actions until it is either confident enough to answer or the turn limit $K_{\max}$ is reached. The complete trajectory is recorded as:

$$\tau = \{(\mathcal{F}_k, x_k, y_k)\}_{k=0}^K.$$

where k indexes the turns starting from the initial turn $k = 0$, and K denotes the terminal turn, with $0 \leq K \leq K_{\max}$.

The complete rollout process is outlined in Algorithm 1.

Algorithm 1 Rollout of Multi-turn Reasoning Trajectory

Input: Long video V , Policy MLLM π_θ , Question x_0 , Input frame set $\mathcal{F}_0$, Maximum turn $K_{\max}$

Output: Final trajectory τ

Initialize: $k \leftarrow 0$, rollout trajectory $\tau \leftarrow (\mathcal{F}_0, x_0)$

```

1: while  $k < K_{\max}$  do
2:   Generate response  $y_k \sim \pi_\theta(\cdot | s_k)$ 
3:    $\tau \leftarrow \tau + y_k$ 
4:    $(reason_k, a_k) \leftarrow Parse(y_k)$ 
5:   if  $a_k$  matches "Retrieval" format then
6:     Extract  $(t_{start}, t_{end})$  from  $a_k$ 
7:      $\mathcal{F}_{k+1} \leftarrow RETRIEVEFRAMES(V, t_{start}, t_{end})$ 
8:      $x_{k+1} \leftarrow x_0$   $\triangleright$  question remains unchanged
9:      $\tau \leftarrow \tau + (\mathcal{F}_{k+1}, x_{k+1})$ 
10:  else if  $a_k$  matches "Answer" format then
11:    break  $\triangleright$  Get final answer
12:  else
13:     $x_k \leftarrow$  "Invalid action. Let me rethink."  $\triangleright$ 
14:    Regenerate response for invalid action
15:     $\tau \leftarrow \tau + (x_k)$ 
16:  end if
17:   $k \leftarrow k + 1$ 
18: end while
19: Collect final trajectory  $\tau$ 

```

While prior studies have applied reinforcement learning to MLLMs for temporal reasoning tasks, they predominantly adopt a single-turn reasoning settings. However, standard RL frameworks for MLLMs struggle with multi-turn optimization due to uniform credit assignment of sparse terminal rewards across turns. This hinders learning nuanced intermediate behaviors that are critical to final success. Furthermore, optimizing solely based on final-task accuracy generally demands extensive training data because terminal supervision is sparse. To address these multi-turn challenges, we introduce a gated bi-level reward system that augments conventional trajectory-level rewards with turn-level rewards. The turn-level rewards encode frame–query relevance, yielding more informative and discriminative signals. As most video question answering datasets provide only QA annotations, we increase data diversity by incorporating a video temporal grounding dataset and curating it to our QA-centric setup. Additionally, observing limited proactive frame retrieval in pretrained MLLMs, we adopt a dynamic exploration-bootstrapping strategy to encourage multi-turn evidence seeking.

3.2. Gated Bi-Level Reward

This section details our fine-grained reward design for RL training. We first describe the computation of the basic bi-level reward. We then present a goal-gated mechanism that prioritizes trajectory-level signals over turn-level ones

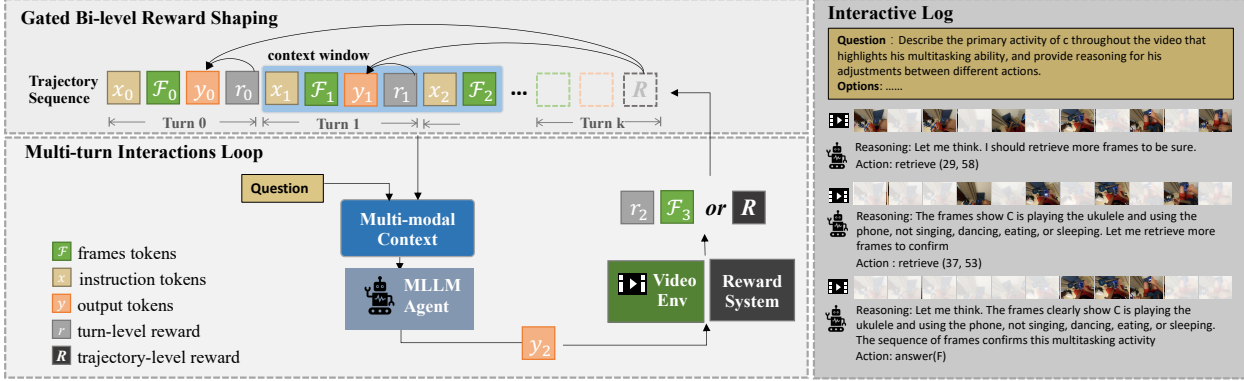


Figure 1. Overview of the proposed Video-MTR framework. *Left*: The lower part shows the multi-turn interaction loop between the MLLM agent and the video environment, while the upper part visualizes the collected trajectory and the gated bi-level reward shaping process during optimization. *Right*: Detailed logs of the agent’s interaction steps across turns.

to align intermediate decisions with the final goal, fostering coherent, goal-oriented multi-turn reasoning.

3.2.1. BI-LEVEL REWARD

This bi-level architecture comprises two complementary components: a trajectory-level reward R_{acc} providing global supervision, and intermediate turn-level rewards to deliver localized feedback within individual turns. The trajectory-level reward R_{acc} is binary, set to 1 if the final answer is correct and 0 otherwise.

R_{fms}^k measures the quality of frame retrieval at the turn level, with a maximum reward of 0.5. This value is set to half of the QA reward, numerically ensuring its role as an auxiliary reward signal. At each intermediate turn k , the relevance of the retrieved frames $\mathcal{F}_k$ to the QA pair is quantified by the IoU with the ground-truth frames $\mathcal{G}$. The IoU score is tracked across turns, and a reward of 0.5 is assigned only if the current retrieval improves upon the best IoU achieved so far; otherwise, a penalty proportional to the IoU drop is applied. This design emphasizes marginal improvements in the retrieved frame set, effectively preventing reward hacking through redundant frame selection while encouraging more efficient evidence gathering.

We also apply a formatting reward of $R_{format}^k = 0.1$ at each turn if the model’s output conforms to the required format. The details of the implementation are provided in Appendix A.2.

3.2.2. GOAL-GATED REWARD SHAPING

To ensure that intermediate actions contribute to the ultimate goal of video understanding, we introduce a goal-gated reward shaping mechanism. In this design, frame-retrieval rewards are granted only when the final answer is correct, ensuring that only retrieval operations leading to successful outcomes are reinforced. This couples retrieval and answering within the policy, rather than optimizing them

in isolation. In our experiments, this setting proved critical. Without such constraints, since frame-retrieval actions can be issued multiple times, the model tended to prioritize optimizing retrieval actions to accumulate positive signals, while neglecting the primary objective of improving video understanding accuracy.

$$R(\tau) = \mathbf{1}_{\{R_{acc} > 0\}} \cdot \sum_{k=0}^{K-1} (R_{fms}^k + R_{format}^k) + R_{acc} + R_{format}^K$$

We aggregate the refined rewards into final reward-annotated trajectories, which then serve as training data for policy optimization.

3.3. Reinforcement Learning

Building upon the formulated reward structure, Video-MTR employs a pure reinforcement learning paradigm to optimize the policy. In this section, we describe how we adapt the standard policy gradient methods to the specific challenges of multi-turn video reasoning.

The standard RL objective function of the trajectory is defined as: $\max_{\pi_{\theta}} \mathbb{E}_{\tau \sim \pi_{\theta}} [R(\tau)]$. We train the policy with Proximal Policy Optimization (PPO) and extend its default formulation to accommodate multi-turn reasoning. The multi-turn interactions trajectory is treated as an entire token sequence $\mathbf{z} = (z_0, z_1, \dots, z_T)$. Instead of relying solely on sparse final-step feedback, the bi-level rewards are applied at every turn boundary and then propagated across all tokens z_t , enabling effective end-to-end learning. Specifically, two discount factors jointly shape the rewards during the calculation of token-level advantages A_t^{GAE} :

- γ_{turn} : a cross-turn discount factor (0.95) applied to the accuracy reward R_{acc} , propagating the final answer signal back to earlier turns. At the boundary of turn

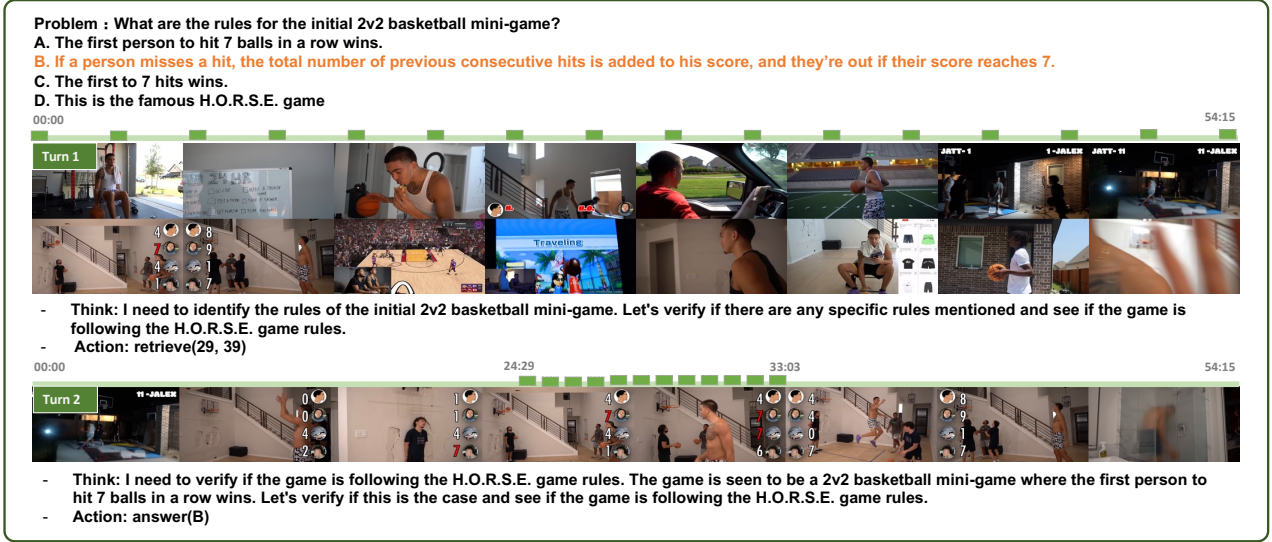


Figure 2. Illustration of Video-MTR’s Multi-turn Reasoning Process, visualizing sampled frames, reasoning process, and model actions per turn. The ground-truth answer is highlighted in orange. The green timeline indicates the positions of sampled frames in the video, reflecting the model’s frame selection strategy at each reasoning turn.

k , the assigned reward is the original frame-retrieval reward of that turn plus a discounted accuracy term:

$$R_{\text{fms}}^k + \gamma_{\text{turn}}^{K-k} R_{\text{acc}}.$$

- γ_{token} : a within-turn discount factor (1.0) that propagates the turn boundary reward to tokens within the same turn.

The computed token-level advantages A_t^{GAE} are then used in the standard PPO surrogate objective, ensuring that the sparse bi-level supervision signals jointly guide policy optimization. In practice, optimizing this objective presents two core challenges: (1) precisely estimating the intermediate frame-retrieval rewards; and (2) shifting a model originally biased toward single-turn reasoning into a multi-turn paradigm. We address these challenges with two strategies: a high-quality data curation pipeline that delivers fine-grained temporal supervision, and an exploration bootstrapping mechanism that incentivizes multi-turn retrieval behavior without the need for warm-up SFT.

Data Curation To obtain precise temporal supervision without incurring additional annotation costs, we repurpose existing video grounding benchmarks, NExT-GQA (Xiao et al., 2024) and QVHighlights (Lei et al., 2021) for our training. NExT-GQA naturally aligns with our requirements, offering QA pairs already equipped with explicit grounding annotations. For QVHighlights which contains only descriptive captions, we further employ a generative adaptation step where GPT-4o (Hurst et al., 2024) converts grounding queries into QA pairs while preserving the original temporal annotations. We then filter for instances where relevant segments are sparse relative to the full video, yielding a highly compact dataset of only 8K samples. By prioritizing su-

pervision fidelity over data volume, this lightweight dataset enables efficient RL post-training that allows Video-MTR to attain competitive performance with significantly less data than large-scale baselines. Further curation details are available in Appendix B.

Exploration Bootstrapping The prevailing post-training paradigms for LLMs and MLLMs typically rely on a two-stage pipeline to mitigate the cold-start problem, a process that often necessitates generated exemplar trajectories. We address this challenge by introducing an adaptive exploration bonus: within each mini-batch, if the agent’s frame-retrieval rate falls below a threshold, each retrieval action receives a small positive reward regardless of relevance; as retrievals become routine, the bonus is automatically disabled. Once the initial retrieval capability is established, the synergy with fine-grained bi-level signals empowers the agent to master complex multi-turn evidence-seeking behaviors, thereby holistically enabling a robust pure RL paradigm without a warm-up SFT.

4. Experiments

4.1. Implementation Details

Video-MTR is built upon the Qwen2.5-VL-7B and trained using the VAGEN (Wang* et al., 2025) framework, which supports multi-turn reinforcement learning. The policy is trained with PPO using a batch size of 32, an actor learning rate of 1×10^{-6} , and a critic learning rate of 1×10^{-5} . Experiments are conducted on a single server equipped with eight NVIDIA A800-80GB GPUs.

Number of Turns We set the maximum number of turns K_{max} to 3, achieving a favorable trade-off between accuracy

Table 1. Performance on mainstream long-video benchmarks. †: results reported in the original paper; *: results from our re-implementation/evaluation under different input settings. Best and second-best per category are **bolded** and underlined, respectively.

Model	Size	Frames	VideoMME	MLVU	LongVideoBench	LVBench	EgoSchema
			Overall(w/o sub.)	Test	Val	Overall	Subset
<i>Proprietary Models or Input Frame Budget: > 256 frames</i>							
GPT-4o (Hurst et al., 2024)	-	0.5 fps / 384	71.9	54.9	66.7	48.9	72.2
Gemini-1.5-Pro (Team et al., 2024)	-	0.5 fps	75.0	-	64.0	33.1	71.1
DrVideo(GPT-4) (Ma et al., 2025)	-	0.2/0.5 fps	51.7	-	-	-	66.4
Qwen2.5-VL-7B† (Bai et al., 2025)	7B	768	65.1	-	56.0	45.3	65.0
VideoLLaMA2 (Cheng et al., 2024)	8×7B	8	47.9	45.6	-	-	53.3
Video-CCAM (Fei et al., 2024)	9B	96	50.3	42.9	43.1	-	-
LongVA (Zhang et al., 2024)	7B	128 / 256	52.6	41.1	47.8	37.9	-
Video-XL (Shu et al., 2025)	7B	128 / 256	55.5	45.6	50.7	-	-
VideoAgent (Wang et al., 2024b)	-	87	56.0	-	-	-	60.2
VideoMemAgent (Fan et al., 2024)	-	72	57.4	-	-	-	62.8
Video-LLaVA (Lin et al., 2023)	7B	8	39.9	30.7	39.1	-	36.8
VideoChat2 (Li et al., 2024b)	7B	16	39.5	30.1	39.3	-	-
LLaVA-OneVision (Li et al., 2024a)	7B	32	58.2	-	56.3	-	60.1
Video-R1 (Feng et al., 2025)	7B	32	59.3	45.4	-	35.9	48.8
Video-R1 (Feng et al., 2025)	7B	64	61.4	47.6	-	38.0	51.8
Qwen2.5-VL-7B* (Bai et al., 2025)	7B	32	53.6	41.6	45.8	30.3	59.4
Qwen2.5-VL-7B* (Bai et al., 2025)	7B	64	58.4	41.8	47.0	33.7	62.6
Qwen2.5-VL-7B* (Bai et al., 2025)	7B	80	59.5	45.2	48.4	33.6	63.5
Video-MTR	7B	32	59.0	48.4	52.3	38.2	62.4
Video-MTR	7B	64	62.2	49.8	54.8	41.8	63.4
Video-MTR (Ours)	7B	80	62.7	50.4	57.1	42.3	68.8

and inference latency. A detailed examination, including quantitative comparisons under varying settings, is reported in the Appendix A.3.

Input Frame Budget Most LVU post-training methods operate with ≤ 128 frames to align with training sequence lengths and manage computation. Given our resource constraints and to emphasize reasoning paradigm rather than raw capacity, we cap the input at 80 frames. Under the same budget, we compare: (i) a single-turn baseline with uniformly sampled frames; and (ii) our multi-turn framework that actively retrieves non-uniform subsets across turns, holding other factors fixed to isolate the effect of multi-turn reasoning. We evaluate budgets of 32, 64 and 80, and results consistently show that distributing frames over multiple retrieval–reasoning steps outperforms the single-turn baselines. Concretely, the first turn uniformly samples half the budget, and each subsequent turn retrieves up to one quarter, ensuring the total never exceeds the frame budget.

4.2. Benchmarks

We select five representative long-form video benchmarks for comprehensive evaluation. Among them, VideoMME (Fu et al., 2025) is one of the most widely used benchmarks for general video understanding. To more closely target the challenges of long-form video reasoning, we further include MLVU (Zhou et al., 2024), LongVideoBench (Wu et al., 2024) and LVBench (Wang et al., 2024a), all featuring significantly extended video durations and complex task de-

signs that rigorously test the capabilities and limitations of current MLLMs. Finally, we include the egocentric benchmark EgoSchema (Mangalam et al., 2023) of first-person human activities to evaluate the model’s generalization across diverse scenarios.

4.3. Performance of Long-form Video Understanding

4.3.1. MAIN RESULTS

We use objective questions across all benchmarks. The main results are summarized in Table 1. For long video understanding, achieving strong performance in prior work typically relies on either ultra-large proprietary models with hundreds of billions of parameters, or processing a substantial number of sampled frames, both of which are highly resource-intensive. For fairness, we report model size and input frame count alongside accuracy. Under comparable parameter and frame scales, Video-MTR shows clear advantages across all benchmarks. This strictly out-of-domain evaluation strongly suggests that Video-MTR learns a generalizable policy across different video domains. Notably, despite its compact 7B parameter size, Video-MTR achieves comparable performance on several challenging datasets, such as MLVU and EgoSchema, when compared to ultra-large proprietary models like GPT-4o and Gemini-1.5-Pro, which have significantly larger parameter counts and process vastly more input frames. Furthermore, compared to its backbone, Video-MTR with **80** input frames already achieves performance comparable to Qwen2.5-VL-7B with

768 frames across most of the datasets, and even outperforms it on EgoSchema (+3.8%) and LongVideoBench (+1.1%). We further analyze Video-MTR’s advantages and summarize key findings below.

Table 2. Comparison of training paradigms, data modalities and volumes. (M)/(S) denote multi-turn and single-turn respectively.

Method	Paradigm	Modalities	Volume
Video-CCAM	SFT	img/vid-text	4.4M
VideoChat2	SFT	img/vid-text	2M
LongVA	SFT	img-text	1.3M
Video-XL	SFT	img/vid-text	257K
Video-R1	SFT+RL (S)	img/vid-text	260K
Video-MTR (Ours)	RL(M)	vid-text	8K

Data-Efficient Training Beyond accuracy, Table 2 compares post-training paradigms and data requirements across representative approaches (see Appendix D for detailed comparisons with all benchmarked baselines). For a strictly fair comparison, we only compare the data used during the fine-tuning stage for LVU. Most counterparts rely on hundreds of thousands to millions of supervised multimodal pairs, whereas Video-MTR is post-trained in a single RL stage with only 8K supervision-rich examples. Despite the drastic reduction in data scale, our model matches or even surpasses methods trained on vastly larger datasets across mainstream long-video benchmarks. To further validate this RL training paradigm, we applied the same procedure to Qwen2.5-VL-3B. Even with this smaller backbone, the model rapidly gained multi-turn reasoning capability, outperforming its original single-turn baseline. Detailed results are provided in Appendix A.4. These findings show that the proposed paradigm is scalable and highly data-efficient. With just one to two training epochs, Video-MTR transforms an open-source MLLM from a single-turn to an iterative reasoner, offering a practical, cost-effective solution for long-video understanding.

Scalability with Frame Budget Given our practical resource constraints, training Video-MTR with hundreds of frames is currently computationally prohibitive. Within this constraint, we implemented Video-MTR across 32, 64, and 80-frame settings, revealing clear scalability with frame budget as shown in Table 1 and Figure 3. We observe two key trends: (i) under matched budgets, Video-MTR consistently outperforms the Qwen2.5-VL-7B backbone, validating the effectiveness of multi-turn reasoning; and (ii) performance exhibits a positive correlation with frame count, showing steady gains as the budget increases. These observations suggest that Video-MTR possesses strong potential to scale further given larger computational resources.

4.3.2. CASE STUDY

Figure 2 illustrates Video-MTR’s multi-turn reasoning on a 54-minute video for a single-detail query hinging on a

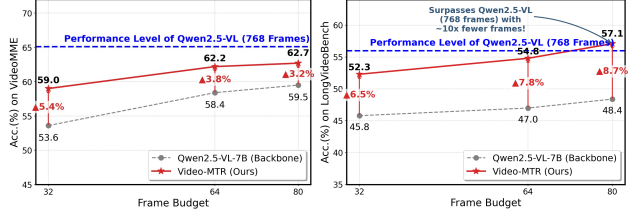


Figure 3. Performance trends across varying frame budgets on LongVideoBench and VideoMME. The plots illustrate Video-MTR’s clear scalability and its consistent superiority over the backbone at matched frame counts.

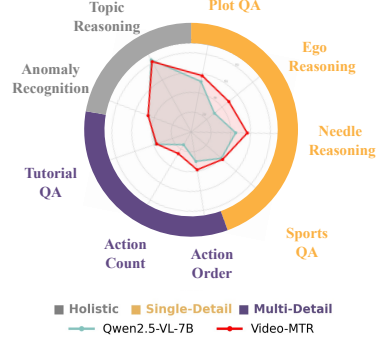


Figure 4. Task Diagnosis on MLVU benchmark.

critical plot point. In Turn 1, frames are uniformly sampled across the entire video. Noting that key evidence is missing, Video-MTR autonomously retrieves densely sampled segments semantically aligned with the query. In Turn 2, it re-examines the refined, query-relevant frames, extracts the required detail, and outputs the correct answer. This case shows how iterative retrieval and focused inspection overcome the limitations of uniform sampling in long videos.

4.4. Ablation Study

We further investigate the contributions of several key components through detailed ablation studies.

4.4.1. ANALYSIS OF THE MULTI-TURN REASONING

We analyze the advantages of the proposed multi-turn reasoning framework over the conventional single-turn

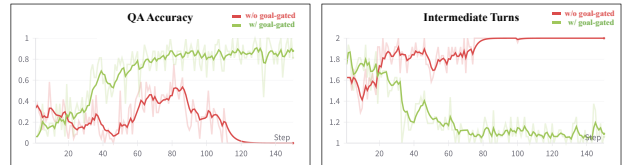


Figure 5. Reward hacking example. The red curve shows in the w/o goal-gated setting, the agent may simply accumulate more turns to increase reward, but with no corresponding gain in QA accuracy.

Table 3. Comparisons of accuracy improvements across video durations.

Model	Frames	VideoMME (w/o sub.)		
		Short	Medium	Long
Qwen2.5-VL-7B	32	65.8	50.3	44.7
Video-MTR	32	70.4 ^{+4.6}	55.6 ^{+5.3}	51.0 ^{+6.3}
Qwen2.5-VL-7B	64	72.1	55.9	47.1
Video-MTR	64	72.8 ^{+0.7}	62.3 ^{+6.4}	51.4 ^{+4.3}
Qwen2.5-VL-7B	80	73.1	56.7	48.3
Video-MTR (Ours)	80	74.8 ^{+1.7}	60.6 ^{+5.9}	52.7 ^{+4.4}

Table 4. Ablation study. The first variant keeps the multi-turn paradigm but removes the bi-level reward. The second variant switches to a single-turn paradigm.

Ablation Setting	VideoMME (w/o sub.)				LVBench
	Short	Medium	Long	Overall	Overall
Ours	74.8	60.6	52.7	62.7	42.3
Ours Multi-turn w/o Bi-Level Reward	69.4	56.2	49.4	58.3	37.7
Ours Single-turn	68.8	54.8	47.9	57.2	35.3

paradigm. Since Video-MTR is built on Qwen2.5-VL-7B, we compare directly against this base model to isolate performance gains. As multi-turn reasoning is expected to be particularly beneficial for complex tasks, we empirically assess its impact across diverse task types and video durations. **(1)Task types.** Using the MLVU benchmark, which categorizes evaluation tasks into three types: holistic tasks (global understanding of the entire video), single-detail tasks (focusing on one critical plot), and multi-detail tasks (requiring reasoning over multiple events), we observe distinct trends in Figure 4. For holistic tasks, typically lower in complexity, the base model achieves up to 72% accuracy, with Video-MTR providing a modest improvement of +3.8%. In contrast, detail-oriented tasks are substantially harder. The base model remains below 40% accuracy, while Video-MTR yields larger gains: +7.5% on single-detail and +8.1% on multi-detail. These results suggest a near-linear relationship between task complexity and the benefits of multi-turn reasoning. **(2)Video durations.** We further examine the impact of duration on VideoMME. We also observe a positive correlation between video length and performance gains. As shown in Table 3, under the 32-frame constraint, Video-MTR achieves accuracy improvements of +4.6% (Short), +5.3% (Medium), and +6.3% (Long) compared to Qwen2.5-VL-7B. Similarly, under the 64/80-frame constraint, the improvements for Medium and Long videos are notably higher than for Short videos.

To ensure a fair comparison, we further post-train Qwen2.5-VL-7B on the same data as Video-MTR. This yields our single-turn baseline, which processes the same number of uniformly sampled frames in a single forward pass. Compared with Video-MTR, it uses the same accuracy-based reward but removes multi-turn instructions from the prompts. Both models use identical optimization hyperparameters. Results for the single-turn baseline are reported in the third row of Table 4. While this single-turn variant yields modest

improvements over Qwen2.5-VL-7B, it falls short when compared to Video-MTR, particularly on complex tasks in LVBench and long-form videos in VideoMME, consistent with our earlier analysis. This performance gap highlights the effectiveness of the multi-turn reasoning paradigm for complex inference.

4.4.2. EFFECTIVENESS OF BI-LEVEL REWARD

We evaluate the bi-level reward design against a multi-turn variant that omits this component, which removes turn-level supervision and relies solely on the final accuracy reward to guide the multi-turn behavior. As shown in Table 4, even with identical prompts and preserved multi-turn behavior, accuracy declines across benchmarks (including a significant 4.6% drop on LVBench). These findings highlight that, without intermediate supervision, relying solely on a final accuracy reward is insufficient to guide the model toward effective temporal localization, thereby limiting its reasoning capability.

4.4.3. NECESSITY OF GOAL-GATED REWARD SHAPING

To assess the effectiveness of our goal-gated reward shaping in mitigating reward hacking, we compare Video-MTR with an ablated variant that removes this mechanism and instead receives unconditioned turn-level rewards. Figure 5 shows the resulting failure mode that emerges early in training: during training, the ablated agent inflates reward by repeatedly retrieving frames with more turns rather than answering correctly. By contrast, the goal-gated model gradually learns to retrieve frames based on actual necessity, thereby optimizing its question-answering capability. These results confirm that goal-gated shaping is crucial for preventing superficial reward exploitation and preserving genuine video understanding capability.

5. Conclusion

We present Video-MTR, a reinforced multi-turn reasoning framework for long-form video understanding. Video-MTR effectively integrates pure reinforcement learning with explicit multi-turn reasoning. At the core of the framework is a gated bi-level reward mechanism, designed to incentivize both relevant frame retrieval and step-by-step reasoning. Extensive experiments on five mainstream benchmarks demonstrate that Video-MTR achieves strong and robust performance across diverse task types and varying temporal lengths. Notably, the framework exhibits excellent temporal scalability, yielding higher gains as video duration increases, highlighting its particular advantage in extra-long video understanding. Future work includes extending the framework to even longer videos and more complex reasoning tasks, pushing the boundaries of long-video understanding.

Impact Statement

This work advances long-video understanding by improving the ability of multimodal models to allocate limited visual evidence and reason over long temporal contexts. The proposed approach may benefit applications such as video retrieval, educational video analysis, assistive video understanding, and efficient long-form media analysis. We do not foresee direct negative societal impacts specific to this work, while noting that future applications of long-video understanding systems should follow appropriate privacy, consent, and safety considerations.

References

- Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Tang, Z., Yuan, L., et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024a.
- Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024b.
- Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.
- Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P. N., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.
- Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., and Li, Q. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.
- Fei, J., Li, D., Deng, Z., Wang, Z., Liu, G., and Wang, H. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024.
- Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., and Yue, X. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.
- Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Lei, J., Berg, T. L., and Bansal, M. Qvhighlights: detecting moments and highlights in videos via natural language queries. In *Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS ’21*, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.
- Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.
- Li, G., Hammoud, H., Itani, H., Khizbullin, D., and Ghanem, B. Camel: Communicative agents for” mind” exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.
- Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.
- Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., and Wang, L. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.
- Li, Y., Wang, C., and Jia, J. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024c.
- Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., and Yuan, L. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.
- Liu, J., Wang, Y., Ma, H., Wu, X., Ma, X., Wei, X., Jiao, J., Wu, E., and Hu, J. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024.

- Ma, Z., Gou, C., Shi, H., Sun, B., Li, S., Rezaatofghi, H., and Cai, J. Drvideo: Document retrieval based long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18936–18946, 2025.
- Mangalam, K., Akshulakov, R., and Malik, J. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.
- Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.
- Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.
- Shu, Y., Liu, Z., Zhang, P., Qin, M., Zhou, J., Liang, Z., Huang, T., and Zhao, B. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26160–26169, 2025.
- Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Wang*, K., Zhang*, P., Wang*, Z., Gao*, Y., Li*, L., Wang, Q., Chen, H., Wan, C., Lu, Y., Yang, Z., Wang, L., Krishna, R., Wu, J., Fei-Fei, L., Choi, Y., and Li, M. Vagen:reinforcing world model reasoning for multi-turn vlm agents, 2025. URL <https://vagen-ai.github.io/>.
- Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024a.
- Wang, X., Zhang, Y., Zohar, O., and Yeung-Levy, S. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024b.
- Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024c.
- Weng, Y., Han, M., He, H., Chang, X., and Zhuang, B. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pp. 453–470. Springer, 2024.
- Wu, H., Li, D., Chen, B., and Li, J. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.
- Wu, P. and Xie, S. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 3(4), 2023.
- Xiao, J., Yao, A., Li, Y., and Chua, T.-S. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.
- Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., and Liu, Z. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.
- Zhao, Y., Misra, I., Krähenbühl, P., and Girdhar, R. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.
- Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., and Yu, X. Deepeyes: Incentivizing” thinking with images” via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.
- Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T., and Liu, Z. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, pp. arXiv–2406, 2024.

A. Training Details

A.1. Prompt Design

This section details our prompt design and provides an illustrative example in Figure 7. To incentivize multi-turn reasoning, we craft an instruction template that guides the MLLM to follow a predefined interaction protocol. The prompt is multimodal: visual tokens corresponding to frames observed in the current turn are inserted immediately after their textual description. We then append a format template that constrains the model’s output to a structured schema. We define two actions per turn: (i) `answer`, which outputs only the single option letter; and (ii) `retrieve`, which outputs `start_frame` and `end_frame`. In each turn, the model is explicitly required to first provide a brief rationale and then emit the action in the specified format.

A.2. Frame Retrieval Protocol

We next describe the frame-retrieval format and implementation. At preprocessing, we uniformly subsample up to M frames from each video to form a candidate pool $\mathcal{F}_{all}$ and index them accordingly; in our implementation we set $M = 128$, which worked well empirically. In the frame budget settings of 32, the agent receives a sparse overview of 16 uniformly spaced frames in the initial turn. In subsequent turns, the agent may issue a retrieval action that selects a temporal interval by outputting `start_frame` and `end_frame` ($\mathcal{F}_{all}$). The environment then returns frames from this interval at an appropriate stride, capped at 8 frames. This procedure allows the model to iteratively focus on key segments by selecting targeted subsets of frames.

A.3. Analysis of Turn Limit

Although multi-turn reasoning improves accuracy through iterative evidence gathering, it requires multiple forward passes, leading to increased inference latency. This creates a fundamental trade-off between efficiency and performance. To quantify it, we conducted controlled experiments under different maximum-turn settings (K_{max}). All experiments are performed with the Qwen2.5-VL-7B backbone, using a fixed total input frame budget of 32 frames to ensure comparability across settings. The model is evaluated on benchmark datasets with identical training and inference conditions, while varying only the maximum number of turns allowed during training. Results in Table 5 show that while additional turns improve accuracy, the gains diminish beyond a certain point, whereas latency grows nearly linearly. Based on this analysis, we set the maximum number of turns K_{max} to 3 and retain the last 2 turns as context, achieving a balanced compromise between accuracy and efficiency.

Table 5. Accuracy on VideoMME and MLVU, latency, and average number of turns used under different maximum-turn settings K_{max} .

Max Turns K_{max}	Avg. Turns Used	Accuracy (%)		Latency (ms)
		VideoMME	MLVU	
1	1.0	54.8	42.6	194.4
2	1.6	57.9	43.1	312.2
3	2.2	59.0	48.4	427.2
5	3.2	60.7	47.4	622.8

A.4. Additional Results on Qwen2.5-VL-3B

To further verify the generality of our end-to-end reinforcement learning training paradigm, we applied the same procedure to Qwen2.5-VL-3B. Despite its smaller capacity compared to Qwen2.5-VL-7B, the model rapidly acquired multi-turn reasoning ability and consistently outperformed its single-turn baseline. These results in Table 6 demonstrate that the proposed framework is not only effective for larger backbones but also generalizes well to lighter models under limited resources.

A.5. Implementation of Exploration Bootstrapping

To address the lack of proactive evidence seeking in early training, we introduce an adaptive exploration bonus that bootstraps multi-turn retrieval. We compute statistics at the mini-batch level (batch size = 32) and use a two-stage schedule. For each mini-batch, if the retrieval rate (fraction of turns issuing a `retrieve` action) falls below a stage-specific threshold, we add

Table 6. Comparison of single-turn and multi-turn settings on Qwen2.5-VL-3B. The multi-turn framework consistently improves accuracy.

Model	Frames	VideoMME (w/o sub.)		MLVU	LVBench	EgoSchema
		Long	Overall	Test	Overall	Subset
3B Models						
Qwen2.5-VL-3B	32	43.6	51.5	41.2	31.2	57.4
Video-MTR (3B)	32	46.8 ^{+3.2}	52.5 ^{+1.0}	42.4 ^{+1.2}	36.1 ^{+4.9}	59.5 ^{+2.1}
Qwen2.5-VL-3B	64	45.9	54.0	43.4	34.7	59.4
Video-MTR (3B)	64	45.4 ^{-0.5}	54.7 ^{+0.7}	47.1 ^{+3.7}	36.7 ^{+2.0}	64.2 ^{+4.8}

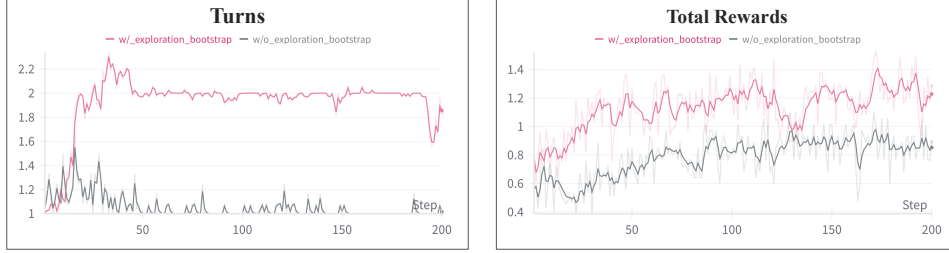


Figure 6. Exploration bootstrapping enables multi-turn behavior. With the bonus (pink), rewards grow as multi-turn retrieval is maintained; without it (gray), the policy stabilizes at single-turn reasoning.

a fixed bonus to every retrieval action in that batch, irrespective of frame relevance.

- Stage I (cold start): threshold = 0.1, bonus = +1.0.
- Stage II (bootstrapping): threshold = 0.5, bonus = +0.5.

Once the retrieval rate remains above the Stage-II threshold for several consecutive mini-batches, the bonus is disabled. As shown in Figure 6, this dynamic shaping reliably kick-starts and sustains multi-turn evidence-seeking behavior under pure RL.

B. Datasets

This section details the construction and statistics of our temporally grounded supervision dataset for reinforcement learning (RL) training. The dataset comprises two components: one curated from a video-understanding dataset NExT-GQA and one adapted from a video temporal grounding dataset QVHighlights:

- NExT-GQA Starting from 10.5K explicit temporal grounding annotations (consolidated into 8.9K QA pairs), we retain instances with a relevant-segment ratio < 0.5 and video duration $> 30s$, yielding $\sim 5K$ high-quality samples.
- QVHighlights We use GPT-4o to convert each original query into a QA pair aligned with its temporal annotations, and apply a two-stage quality filter: (i) discriminative-adequacy screening; and (ii) relevant-segment ratio < 0.5 and video duration $> 30s$, resulting in $\sim 3K$ QA-grounded samples.

In total, we obtain 8K training instances that are compact yet supervision-dense. Table 7 reports per-source composition and retained counts at each step to facilitate reproduction and extension. Figure 8 illustrates the GPT-4o prompt design for rewriting and provides before/after examples.

Table 7. Dataset composition and filtering statistics. Counts denote thousands of samples. NExT-GQA is directly used as QA pairs.

Source	Pre Filter	QA Converted	Post Filter
NExT-GQA	8.9K	-	4.9K
QVHighlights	7.2K	3.5K	3.0K

System: conversation between User and Assistant. The user asks a question, and the Assistant solves it. You are an assistant in analyzing videos. You will be given a video and a question. Goal: Answer the question correctly with no more than 3 turns.

User: Turn 1. Now you are given 16 selected frames from the video, with frame_idx_list: [0 4 8 12 16 21 25 29 33 37 42 46 50 54 58 63].
 Frames: frame_idx:0, <image_0>, frame_idx:4, <image_4>, frame_idx:8, <image_8>, ..., frame_idx:54, <image_54>, frame_idx:58, <image_58>, frame_idx:63, <image_63>
 Answer the following problem based on the frames: What is the event in the video? Options: ['A. Women's synchronized diving', 'B. Individual diving', 'C. Swimming' 'D. Men's synchronized diving', 'E. Swimming relay race', 'F. Synchronized swimming'] You can choose to retrieve more frames or provide your answer.
 Format Template:
 <think>...</think><answer>...</answer> or <think>...</think><retrieve>...</retrieve>
 Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> and </think> tags. If you have enough information, Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.
 If you lack some information, think about the most relevant frame index range of the information you need, then you can retrieve dense frames in the range by sending a retrieve request by <retrieve> start_frame, end_frame </retrieve>.
 IMPORTANT: start_frame and end_frame must be integers smaller than 63.

Video-MTR: <think>I need to identify the event and verify if it is men's synchronized diving. Let's look for any indication that this is a synchronized diving event and verify if it is men's synchronized diving.</think><retrieve>12,33</retrieve>

User: Turn 2. Now you are given 8 selected frames from the video, with frame_idx_list: [12 15 18 21 24 27 30 33].
 Frames: frame_idx:12, <image_12>, frame_idx:15, <image_15>, frame_idx:27, <image_27>, ..., frame_idx:30, <image_30>, frame_idx:33, <image_33>,
 Answer the following problem based on the frames: What is the event in the video?
 ...
 If you lack some information, think about the most relevant frame index range of the information you need, then you can retrieve dense frames in the range by sending a retrieve request by <retrieve> start_frame, end_frame </retrieve>.

Video-MTR: <think>I need to verify if this is a synchronized diving event and check if it is men's synchronized diving. The frames show two divers performing a synchronized dive, which indicates a synchronized diving event. </think><answer>D</answer>

Figure 7. An example of Video-MTR with interactions of 2 turns. <image_n> will be replaced with the visual tokens of the specific frames.

C. Case Studies

We present additional case studies drawn from three evaluation benchmarks—VideoMME (Fu et al., 2025), MLVU (Zhou et al., 2024), and EgoSchema (Mangalam et al., 2023) to give a comprehensive picture of Video-MTR’s multi-round reasoning process; these examples include both successes and failures.

C.1. Successful Cases

From each dataset we randomly selected one correctly solved example. As illustrated in Figure 9, all three examples exhibit a consistent evidence-seeking pattern with the following characteristics: (i) an initial global pass over the video produces a tentative hypothesis that roughly answers the question; (ii) the model then proposes a targeted temporal segment for closer inspection to obtain discriminative evidence; and (iii) after observing this segment, the model updates or confirms the hypothesis and outputs the final answer.

Case A (role identification). The query asks for the identities of two people. After the first pass, Video-MTR hypothesizes that the pair may be a teacher and a student based on coarse contextual cues from the full video. It then narrows attention to their interaction segment for verification. In that focused clip, the person in a white shirt is seen giving instructions, and the standing man in a black shirt follows the instructions and plays the instrument. This instructional exchange provides role-asymmetric signals: directive speech acts, demonstrative gestures, and action-response ordering, yielding temporally grounded, discriminative evidence that confirms the teacher-student hypothesis.

Case B (event recognition). The question asks which event is shown, with candidates including individual/synchronized diving, swimming, relay, and synchronized swimming. After a global pass, Video-MTR sets a verification subgoal: to confirm synchronized diving—and proposes a discriminative interval for inspection. Focusing on this clip, the model observes two divers executing the same dive with mirrored body alignment, thereby ruling out individual diving and all

swimming events. The model confirms the hypothesis and outputs (D) Men’s synchronized diving.

Case C (goal reasoning). The query seeks a concise account of C’s objective and decisions. After a first pass, Video-MTR hypothesizes that C is choosing what to wear and proposes a targeted interval for verification. In this segment, C looks at various clothes, picks them up, and appears to be deciding what to wear, with no behaviors indicative of folding, packing, ironing, or washing. The model confirms the hypothesis and outputs (C) deciding what clothes to wear.

C.2. Error Analysis and Limitations

We also examine failure cases to diagnose error sources and outline potential remedies. Two representative cases, one involving multi-detail reasoning and the other requiring fine-grained perception are illustrated in Figure 10.

Case A (Action Order). This example falls under the action-order category, a multi-detail task requiring inspection of multiple, disjoint segments. In Rounds 1–2 the sampled frames do not cover all events referenced by the options; nevertheless, the model commits to a prediction, exhibiting hallucination under insufficient evidence. More retrieval rounds are needed to reach a reliable decision. A likely cause is a training-distribution bias: in our data, one to three rounds typically suffice to locate relevant frames and answer correctly, which encourages early stopping even when evidence is incomplete. A straightforward remedy is to expand the curriculum with more sequences that demand four to six retrieval rounds and span widely separated events, prompting the model to keep searching until each candidate answer has been either supported or ruled out.

Case B (Fine-grained Procedural Reasoning). This task requires interpreting micro-actions (e.g., dipping or swishing in a cup versus mixing on a palette) and linking them causally to paint subtlety. Under the current frame-processing pipeline, which must accommodate long temporal sequences, the spatial resolution is kept relatively coarse; as a result, these discriminative cues are likely to appear heavily blurred. To address this limitation, the retrieval-and-reasoning loop at the frame-selection level could be augmented with a hierarchical temporal-to-spatial reasoning mechanism: once a relevant frame segment is identified, the system would crop the corresponding frames and re-analyse high-resolution regions of interest, enabling direct verification of micro-movements before any answer is produced.

These failure cases reveal several structural weaknesses that limit the current version of Video-MTR in complex scenarios. Together, these issues indicate that Video-MTR needs deeper temporal search policies, hierarchical zoom-in vision modules to handle multi-event reasoning and fine-grained perception reliably.

D. More Comparisons

To give a more comprehensive comparison: in addition to the original parameter size and frame budget Table 1, we now also summarize, for each baseline, its backbone LLM and post-training data scale. Regarding the backbone comparison, Table 8 shows that our setting is fair across different implementation choices: Video-MTR shares the exact same 7B Qwen2.5-VL-7B backbone with Video-R1 and uses the same 7B Qwen2 family as LongVA and Video-XL, yet achieves superior performance while being trained on significantly less data (only an 8K long-video QA corpus). In contrast, many strong baselines rely on proprietary GPT-4/Gemini backbones or web-scale multimodal data. For the post-training data, to ensure a fair comparison, we compare only the data used in the post-training stage (instruction tuning or RL), rather than the full pre-training corpora. For this reason, we do not list the massive datasets used to build GPT-4/Gemini or the Qwen2.5-VL-7B backbone itself. Most counterparts rely on hundreds of thousands to millions of supervised multimodal pairs, whereas Video-MTR is post-trained in a single RL stage with only 8K supervision-rich examples, clearly highlights the strong data efficiency of our framework.

E. Future Work

Although Video-MTR demonstrates strong reasoning performance on current long-form benchmarks, ample room for improvement remains when tackling more challenging queries and much longer videos. Future work should therefore advance the multi-round framework on two fronts: (i) lengthen the dialogue loop to support deeper chains of reasoning that solve multi-stage tasks, and (ii) incorporate a hierarchical temporal-to-spatial strategy that begins with coarse video sweeps and adaptively zooms into high-resolution frame crops, thereby securing reliable evidence at both event-level and micro-action scales.

Table 8. Summary of compared baseline models, their backbones, frame budgets, and post-training data scale.

Model	Params	Backbone (LLM)	Post-train Data	Frames / fps
GPT-4o (Hurst et al., 2024)	–	GPT-4o (proprietary)	–	0.5 fps / 384
Gemini-1.5-Pro (Team et al., 2024)	–	Gemini (proprietary)	–	0.5 fps
DrVideo (GPT-4) (Ma et al., 2025)	–	GPT-4 (proprietary)	–	0.2 / 0.5 fps
Qwen2.5-VL-7B [†] (Bai et al., 2025)	7B	Qwen2.5-VL-7B	–	768
VideoLLaMA2 (Cheng et al., 2024)	8×7B	Mixtral-8x7B-Instruct	1.35M	8
Video-CCAM (Fei et al., 2024)	9B	Yi-1.5-9B-Chat	4.4M	96
LongVA (Zhang et al., 2024)	7B	Qwen2-7B-Instruct	760K	128 / 256
Video-XL (Shu et al., 2025)	7B	Qwen2-7B	257K	128 / 256
VideoAgent (Wang et al., 2024b)	–	GPT-4 (proprietary)	–	87
VideoMemAgent (Fan et al., 2024)	–	GPT-4 (proprietary)	–	72
Video-LLaVA (Lin et al., 2023)	7B	Vicuna-7B-v1.5	765K	8
VideoChat2 (Li et al., 2024b)	7B	Vicuna-7B-v0	2.0M	16
LLaVA-OneVision (Li et al., 2024a)	7B	Qwen-2-7B	4.8M	32
Video-R1 (Feng et al., 2025)	7B	Qwen2.5-VL-7B	260K	32 / 64
Video-MTR (Ours)	7B	Qwen2.5-VL-7B	8K	32 / 64 / 80

Goal: Given a declarative sentence to serve as a query for retrieving relevant video segments, generate a multiple choice question.
Follow these rules:

1. Suitability Check: Return False if the sentence is too short or Lacks distinctive details for discriminative options. Else, return True and proceed.
2. Question Format: Use one of these interrogatives: Where, How, Why, What, When, Who
3. Options: Derive one correct answer and three incorrect answers from the sentence.
4. Answer: The correct answer to the question.

Format

```
{
  "suitable": bool, # True/False
  "question": str, # MCQ text (if suitable)
  "options": list,
  "answer": str # Correct option
}
```

Examples

- Sentence: A man in white shirt discusses the right to have and carry firearms.
- Output: {
 "suitable": True
 "question": "What is the man in a white shirt discussing?"
 "options": ["A. The war happens in Europe.", "B. The recent massacre in the US.", "C. The right to have and carry firearms.", "D. The recent crime in the US."]
 "answer": C
}

- Sentence: Woman holds her shopping bags.
- Output: {
 "suitable": False
 "question": ""
 "options": ""
 "answer": ""
}

QA Converted Examples

```
"- query": "Asian chef with dyed pink hair cooks food."
"- question": "What is the Asian chef with dyed pink hair doing?"
"- options": ["A. Preparing ingredients", "B. Serving customers", "C. Cleaning the kitchen", "D. Cooking food"],
"- answer": "D"

"- query": "Two people from the same show interview a man at his house."
"- question": "Where do two people from the same show interview a man?"
"- options": ["A. At his house", "B. In a studio", "C. Outside", "D. In an office"]
"- answer": "A"
```

Figure 8. The GPT-4o prompt template for converting declarative queries into multiple-choice QA pairs with suitability check, options generation, and converted QA examples.

VideoMME

A Problem : According to what is shown in the video, what might be the relationship between the old man in a white shirt and the standing man in a black shirt?

A. Father and son .B. Grandfather and grandson. C. Employer and employee. **D. Teacher and student.**

00:00 31:25

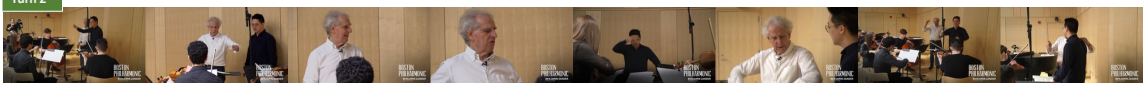
Turn 1



- Think: I need to identify the relationship between the old man in a white shirt and the standing man in a black shirt. Let's verify if the old man in a white shirt is the teacher and the standing man in a black shirt is the student
- Action: retrieve(42,58)

00:00 20:37 28:28 31:25

Turn 2



- Think: I need to verify if the old man in a white shirt is the teacher and the standing man in a black shirt is the student. The old man in a white shirt is seen giving instructions and the standing man in a black shirt is seen following the instructions and playing the instrument. Let's verify if this is happening multiple times and see if the old man in a white shirt is the teacher and the standing man in a black shirt is the student
- Action: answer(D)

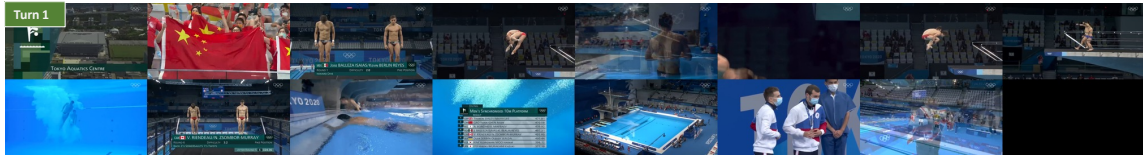
MLVU

B Problem : What is the event in the video?

A. Women's synchronized diving B. Individual diving.
C. Swimming **D. Men's synchronized diving.**
E. Swimming relay race F. Synchronized swim

00:00 96:27

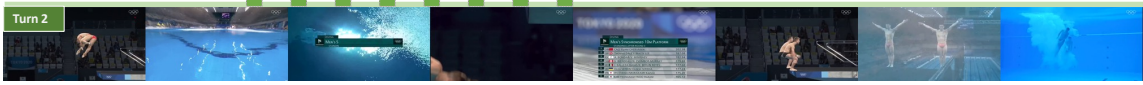
Turn 1



- Think: I need to identify the event and verify if it is men's synchronized diving. Let's look for any indication that this is a synchronized diving event and verify if it is men's synchronized diving.
- Action: retrieve(12,33)

00:00 18:05 49:44 96:27

Turn 2



- Think: I need to verify if this is a synchronized diving event and check if it is men's synchronized diving. The frames show two divers performing a synchronized dive, which indicates a synchronized diving event.
- Action: answer(D)

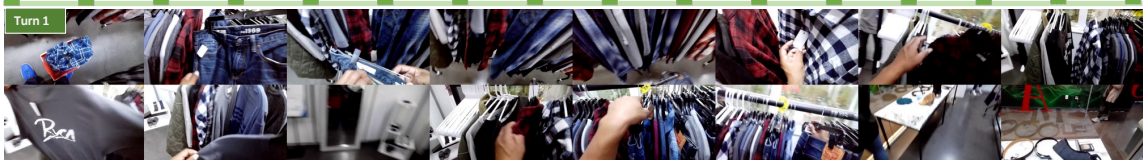
EgoSchema

C Problem : Keeping in mind the ability to compress information, what would be a concise explanation for the main objective and key decisions behind c's actions during the video?

A. C is diligently attempting to clean and wash the clothes thoroughly.
B. Currently, c is attempting to carefully fold and organize the clothes.
C. C is trying to decide what clothes to wear.
D. C is trying to iron the clothes.
E. E. Currently, c is attempting to carefully pack the clothes into luggage.

00:00 03:00

Turn 1



- Think: I need to identify the main objective and see if C is trying to decide what clothes to wear, fold, organize, or pack. Let's verify if C is seen trying to decide what clothes to wear, if C is seen folding, organizing, or packing the clothes.
- Action: retrieve(29,58)

00:00 01:22 02:43 03:00

Turn 2



- Think: I need to verify if C is seen trying to decide what clothes to wear, if C is seen folding, organizing, or packing the clothes. The frames show C looking at various clothes, picking them up, and seems to be deciding what to wear. Let's verify if C is seen trying to decide what clothes to wear, if C is seen folding, organizing, or packing the clothes
- Action: answer(C)

Figure 9. Representative success cases from (A) VideoMME, (B) MLVU, and (C) EgoSchema. The ground-truth answer is highlighted in orange. The green timeline indicates the positions of sampled frames in the video.



Figure 10. Representative failure cases: (A) action-order reasoning error and (B) fine-grained procedural misrecognition. The ground-truth answer is highlighted in orange. The green timeline indicates the positions of sampled frames in the video.