VIDEO-RTS: Rethinking Reinforcement Learning and Test-Time Scaling for Efficient and Enhanced Video Reasoning

Ziyang Wang[1,*] **Jaehong Yoon**[1,2,*] **Shoubin Yu**[1]
Md Mohaiminul Islam[1] **Gedas Bertasius**[1] **Mohit Bansal**[1]

[1]UNC Chapel Hill [2]Nanyang Technological University

https://sites.google.com/cs.unc.edu/videorts2025/

Abstract

Despite advances in reinforcement learning (RL)-based video reasoning with large language models (LLMs), data collection and fine-tuning remain significant challenges. These methods often rely on large-scale supervised fine-tuning (SFT) with extensive video data and long Chain-of-Thought (CoT) annotations, making them costly and hard to scale. To address this, we present VIDEO-RTS, a new approach to improve video reasoning capability with drastically improved data efficiency by combining data-efficient RL with a video-adaptive test-time scaling (TTS) strategy. Building on observations about the data scaling, we skip the resource-intensive SFT step and employ efficient pure-RL training with output-based rewards, requiring no additional annotations or extensive fine-tuning. Furthermore, to utilize computational resources more efficiently, we introduce a sparse-to-dense video TTS strategy that improves inference by iteratively adding frames based on output consistency. We validate our approach on multiple video reasoning benchmarks, showing that VIDEO-RTS surpasses existing video reasoning models by 2.4% in accuracy using only 3.6% training samples. Specifically, VIDEO-RTS achieves a 4.2% improvement on Video-Holmes, a recent and challenging video reasoning benchmark. Notably, our pure RL training and adaptive video TTS offer complementary strengths, enabling VIDEO-RTS's strong reasoning performance.

1 Introduction

Large language models (LLMs) have demonstrated strong problem-solving abilities across diverse domains, enabled by techniques such as Chain-of-Thought (CoT) reasoning (Wang et al., 2022; Yao et al., 2023) and multi-agent collaboration (Talebirad and Nadiri, 2023; Chen et al., 2024b). Building on advances in the language domain, several approaches (Liu et al., 2025; Fei et al., 2024; Feng et al., 2025; Sun et al., 2025; Li et al., 2025a) have recently extended them to improve video reasoning capabilities. However, these methods demand high computational costs and lower training efficiency, typically following an extensive two-stage recipe: (i) supervised fine-tuning (SFT) on reasoning-focused prompts with step-by-step chain-of-thought annotations, followed by (ii) large-scale reinforcement learning using rewards over massive collections of video question-answering data. This pipeline poses substantial computational overhead, particularly in generating long CoT data for video corpus, which limits its scalability for complex, long-term video reasoning tasks.

To overcome these limitations and enable efficient video reasoning, we propose VIDEO-RTS, a novel approach that integrates data-efficient reinforcement learning with video-adaptive test-time scaling strategies, significantly enhancing reasoning performance while maintaining efficiency. In training, unlike the existing approaches (Wang et al., 2025a; Feng et al., 2025), which rely on large-scale supervised fine-tuning (SFT) data with long CoT annotation, we skip the data generation step and directly utilize pure RL training on simple video question-answering (QA) data. Specifically, we adapt the outcome-supervised RL (group relative preference optimization, GRPO (Shao et al., 2024)), motivated by DeepSeek-R1-Zero (DeepSeek-AI, 2025), for its simplicity and effectiveness in aligning model outputs with answer correctness. With merely 6K video-question pairs for RL, our approach matches the performance of the existing SFT+RL framework (Video-R1 (Feng et al., 2025)), which relies on 165K SFT examples plus 4K RL examples, underscoring the effectiveness and training efficiency of VIDEO-RTS.

Furthermore, as illustrated in Fig. 3, scaling to even more video QA samples only brings marginal

[*]Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28115–28129
November 4-9, 2025

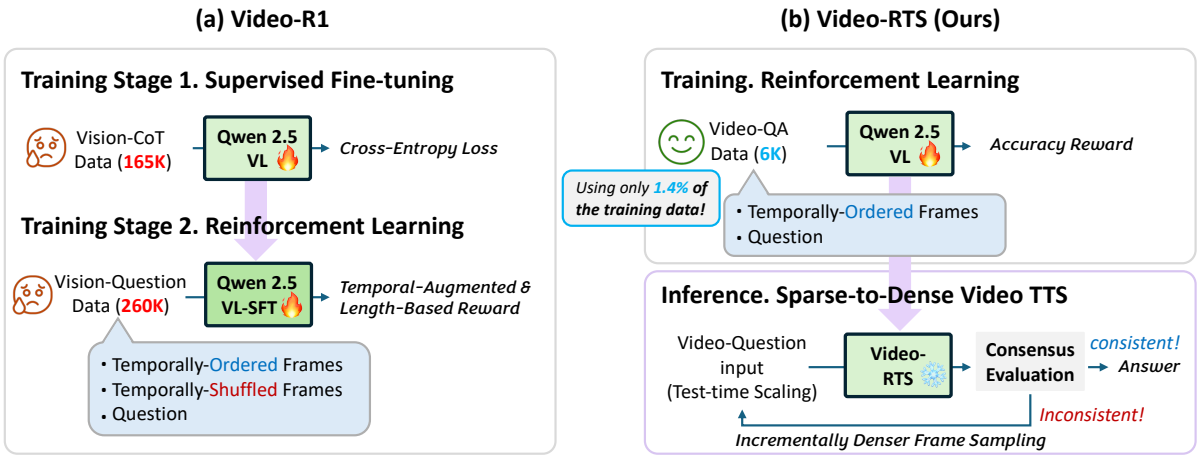

Figure 1: **Training and inference recipe comparison between Video-R1 (Feng et al., 2025) and our VIDEO-RTS.** While (a) Video-R1 uses a two-stage pipeline with SFT and RL, (b) VIDEO-RTS adapts a pure-RL approach with output-based rewards for better data efficiency. We further enhance the reasoning of VIDEO-RTS by proposing dynamic sparse-to-dense video test-time scaling. The format reward is omitted, as both models use it.

improvements, suggesting that the RL training saturates quickly on video reasoning data. This matches the recent findings in the language domain (Wang et al., 2025c) that very few RL training samples could bring great improvement on reasoning tasks. Thus, inspired by the test-time scaling works (Wang et al., 2022; Yao et al., 2023; Snell et al., 2024) in the language community, we aim to enhance the video reasoning capability at the inference stage to better allocate the computational resources. To the best of our knowledge, this is the first study to systematically explore the combination of reinforcement learning and test-time inference strategies for improving video reasoning capability.

To better allocate the excessive training computation, we propose a sparse-to-dense test-time scaling mechanism specifically designed for video reasoning. Specifically, VIDEO-RTS adaptively selects the appropriate temporal context based on output consistency by iteratively adding more frames at the inference stage. Taking advantage of the pure-RL training, the model is able to generate a diverse deep reasoning process given the challenging video query, which allows us to utilize a self-consistency check to decide whether the model obtains sufficient temporal context. The combination of efficient training and adaptive inference enables the model to adapt its computational effort based on the complexity of each input query, producing accurate responses while using only the necessary amount of resources.

We evaluate VIDEO-RTS on the five popular video reasoning benchmarks, including Video-Holmes (Cheng et al., 2025), VideoMMMU (Hu et al., 2025), MMVU (Zhao et al., 2025), VideoMME (Fu et al., 2024a) and LongVideoBench (Wu et al., 2024). Results show that across all benchmarks, compared to the recent Video-R1 model (Feng et al., 2025), which trained on $169K$ samples, VIDEO-RTS, trained with only $6K$ samples (i.e., 96.4% fewer samples), outperforms by 2.4% in average accuracy while using fewer frames during inference. Specifically, on Video-Holmes, the recently proposed complex video reasoning benchmark, VIDEO-RTS outperforms Video-R1 by 4.2%, demonstrating the efficiency and effectiveness of our framework. Furthermore, we find that our pure RL training and sparse-to-dense video test-time scaling are complementary: RL enhances the MLLM's reasoning capabilities, while VIDEO-RTS leverages diverse reasoning strategies to adaptively select the optimal temporal context (i.e., number of frames) for each video query.

2 Related Works

Long Video Understanding. The rise of video understanding models has expanded from short videos to long-video tasks such as classification (Wu and Krahenbuhl, 2021; Mohaiminul Islam and Bertasius, 2022; Islam et al., 2023), captioning (Zhou et al., 2018; Krishna et al., 2017; Islam et al., 2024), and question answering (Fu et al., 2024a; Zhou et al., 2024; Wu et al., 2024). The

emergence of multimodal large language models (MLLMs) (Bai et al., 2025a; Zhang et al., 2024b; Li et al., 2024; Wang et al., 2024b; Bai et al., 2025b; Zhang et al., 2024a; Islam et al., 2025; Wei et al., 2025) has further propelled research in long-video understanding. However, most existing MLLMs focus solely on generating answers without providing rationale or reasoning. We address this gap by proposing a new approach that enables MLLMs to generate both answers and step-by-step reasoning through data-efficient pure-RL training and a video-adaptive test-time scaling mechanism, enhancing interpretability and reducing overfitting.

Visual CoT Reasoning with RL. Inspired by the reasoning capabilities demonstrated by large language models (LLMs) in NLP (DeepSeek-AI, 2025; OpenAI, 2024b), recent efforts have focused on enhancing the reasoning abilities of MLLMs in visual data. Early works targeted image-based reasoning, often using hand-crafted CoT structures (Xu et al., 2024; Thawakar et al., 2025) and modality bridging techniques (Yang et al., 2025; Huang et al., 2025). On the other hand, in the video domain, some approaches focused on temporal grounding (Wang et al., 2024a), while others employed manual reasoning pipelines for general video understanding (Liu et al., 2025; Fei et al., 2024). Lastly, several recent works (Meng et al., 2025; Wang et al., 2025b; Sun et al., 2025; Zhang et al., 2025b; Dang et al., 2025; Li et al., 2025b) have employed Reinforcement Learning (RL) (Kaelbling et al., 1996) strategies such as DPO(Rafailov et al., 2023) and GRPO (Shao et al., 2024) for enhancing MLLM reasoning capabilities. However, the leading methods (Wang et al., 2025a; Feng et al., 2025; Tian et al., 2025) often rely on a costly SFT stage with large amounts of long CoT data. Instead, we use a GRPO-based RL strategy without requiring any SFT data or expensive temporal ordering supervision, enabling data- and compute-efficient training of the video reasoning model.

Test-Time Scaling (TTS). TTS (Snell et al., 2024) refers to the strategic allocation of increased computational resources during inference to facilitate more deliberate, step-by-step reasoning rather than rapid, heuristic processing. While the Chain-of-Thoughts (CoT) framework initially proposed augmenting computational budget during inference to enhance reasoning capabilities, more sophisticated methods have since emerged for

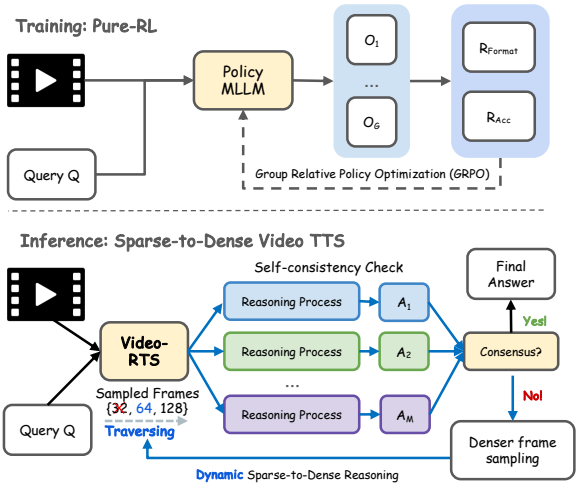

Figure 2: **The overview of VIDEO-RTS.** The training phase (Top) adapts GRPO-based RL to optimize the MLLM with outcome and accuracy rewards. In inference (Bottom), VIDEO-RTS conducts dynamic sparse-to-dense reasoning by traversing sampled frames for generating rationales. If answers are in consensus, it returns an answer; otherwise, it samples denser frames.

language tasks, including self-consistency (Wang et al., 2022), weighted voting (Wan et al., 2025), Tree-of-Thoughts (Yao et al., 2023) and self-reflection (Shinn et al., 2023). However, we argue that text-centric approaches are sub-optimal on complex video understanding tasks, as they overlook the unique characteristics of videos and the varying levels of reasoning granularity required by different queries. Thus, we propose a novel video-adaptive test-time scaling mechanism tailored for the challenges of efficient long-range video reasoning task. Specifically, VIDEO-RTS dynamically allocate inference budget by a sparse-to-dense manner based on the model output consistency.

3 VIDEO-RTS

We propose Video-RTS, a resource-efficient RL and test-time scaling framework for video reasoning. We begin by introducing outcome-supervised RL method, which serves as our base RL algorithm Sec. 3.1. In Sec. 3.2, we define the problem statement and challenges of the video reasoning problem. Next, we propose an efficient reinforcement fine-tuning strategy that leverages simple video QA data without costly chain-of-thought annotations or temporal labels in Sec. 3.3. Finally, we introduce a video-specific test-time scaling mechanism that adaptively adjusts computation, further enhancing performance in Sec. 3.4.

3.1 Preliminary: Group Relative Policy Optimization (GRPO)

Recently, DeepSeek-R1 (DeepSeek-AI, 2025) achieves state-of-the-art performance on multiple language reasoning benchmarks with a newly suggested reinforcement learning (RL) approach. As a key step of the framework, DeepSeek-R1 utilizes Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as the core algorithm to conduct reasoning-oriented RL. Compared to the canonical DPO (Rafailov et al., 2024), GRPO eliminates the need for a value model by estimating baselines from group-level scores. Directly comparing groups of candidate responses removes reliance on a critic model and substantially reduces training costs. Given the input question, GRPO first generates G distinct candidate responses $\{O_1, \ldots, O_G\}$ through different sampling settings from the old policy $\pi_{\theta_{old}}$. The model serves as the reward function to get the corresponding scores $\{R_1, \ldots, R_G\}$. Then the model computes the mean and standard deviation of the candidate's score for normalization and determines the quality of these responses:

$$S_i = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^N)}{\text{std}(\{R_i\}_{i=1}^N)}, \tag{1}$$

where S_i represents the relative quality score of the i-th answer candidates. Given the reasoning question $q \sim P(Q)$, GRPO optimizes the policy model π_θ by maximizing the following objective:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \\ \frac{1}{G}\sum_{i=1}^{G}\Bigg[\min\Bigg(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}S_i, \\ \text{clip}\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\epsilon, 1+\epsilon\right)S_i\Bigg) \\ -\beta\,\mathbb{D}_{\text{KL}}\Big(\pi_\theta\Big\|\pi_{\text{ref}}\Big)\Bigg].\end{aligned} \tag{2}$$

To prevent the updated policy π_θ, parameterized by θ, from drifting too far from the reference model π_{ref}, GRPO incorporates a KL-divergence term $\mathbb{D}_{\text{KL}}$ that penalizes per-token deviations. In this work, we adopt GRPO as our reinforcement learning algorithm to efficiently enhance video reasoning capabilities.

3.2 Problem Statement and Challenges

We formulate the video reasoning task as a video question-answering problem, where given a video input V and a reasoning question Q, the video reasoning model f_θ is designed to generate its predicted answer $\widehat{A}$. Following the standard practice in the recent video reasoning benchmarks (Hu et al., 2025; Fu et al., 2024b), we focus on the multiple-choice question-answering format (MCQA), which also adds answer options A_o as input and requires the model f_θ to choose between the given answer candidates. Concretely, the video reasoning process could be formulated as:

$$\widehat{A} = f_\theta\big(V, Q, A_o\big). \tag{3}$$

Recently, a few notable works (Feng et al., 2025; Sun et al., 2025) show the strong potential of combining supervised fine-tuning (SFT) and reinforcement learning (RL) for addressing video reasoning problems. These methods typically follow a two-stage pipeline: (1) SFT with long Chain-of-Thought (CoT) video QA data, and (2) reasoning-focused RL on video QA data. Despite their effectiveness, several challenges remain: (i) **data inefficiency**: reliance on large-scale video-question or CoT datasets hinders scalability to complex video tasks (e.g., Video-R1 utilize $165K$ SFT data and $24K$ RL data), (ii) **computational inefficiency during RL**: training with dense video-text pairs and complex reward designs is resource-intensive (e.g., temporal GRPO (Feng et al., 2025)), (iii) **limited inference-time adaptability**: current models lack mechanisms to scale computation dynamically based on query complexity.

To address these challenges, we develop VIDEO-RTS, a data-efficient, yet strong video reasoning model that introduces an advanced training recipe along with a consensus-based hierarchical voting strategy for inference.

3.3 Resource-Efficient RL for Video Reasoning

We introduce the proposed RL training strategy of VIDEO-RTS that overcomes the limitations of machine-generated CoT data and the overhead of supervised fine-tuning. The pioneering video reasoning approach, Video R1 (Feng et al., 2025), leverages an open-source MLLM (Qwen-2.5-VL-72B (Bai et al., 2025a)) to generate reasoning chains over $165K$ video QA examples for supervised fine-tuning. Generating large-scale, long-form reasoning chains is time-consuming, and the quality of the resulting SFT data remains uncertain, as the MLLM shows a significant performance gap compared to human experts and lacks fine-tuning on video-specific CoT reasoning formats.

Motivated by the success of DeepSeek-R1-Zero (DeepSeek-AI, 2025), we revisit the standard training pipeline and propose to bypass the costly SFT stage, instead exploring a pure reinforcement learning approach for video QA with minimal training overhead. To equip the video reasoning capabilities in recent powerful image reasoning MLLMs, we apply outcome-supervised RL (i.e., GRPO) directly on video QA data, using a simple reward function based solely on answer correctness, without relying on any additional verifier. The details of each component are described below.

Backbone MLLM. As demonstrated in DeepSeek-R1 (DeepSeek-AI, 2025), an important prerequisite for effective outcome-supervised RL training is the cold-start supervised fine-tuning (SFT) stage, which enhances the model's basic reasoning ability. Prior works on frame-based video understanding (Buch et al., 2022; Lei et al., 2022) have shown that models extensively trained on image data can achieve strong performance on video tasks. Based on this insight, we use an MLLM (e.g., Qwen-2.5-VL (Bai et al., 2025a)) trained on image reasoning data as a strong cold-started model for outcome-supervised RL training on video data.

Reward Design. Inspired by DeepSeek-R1-zero (DeepSeek-AI, 2025), we propose that directly optimizing for outcome-based rewards, rather than relying on step-by-step supervision as in video-SALMONN-o1 (Sun et al., 2025), can further improve reasoning capabilities while reducing the need for costly intermediate CoT data. Moreover, acquiring detailed supervision for intermediate reasoning steps often demands complex verifier designs. To address this, we adopt an efficient reward design that directly optimizes the model's final output O. Specifically, we introduce two types of rewards: *format reward* and *accuracy reward*, to fine-tune the backbone MLLM and induce explicit CoT reasoning ability on complex video tasks.

- First, we apply a format reward R_{format} that encourages the model to generate its reasoning process between '<think>' and '</think>' tags before generating the answer prediction. This reward helps the model to have an explicit logical reasoning step in response to the video and text query before producing the final answer.
- Next, we introduce an accuracy reward R_{acc}, which incentivizes the model to produce correct answers following its reasoning process. We formulate the training task as a multiple-choice QA problem, enabling a straightforward definition of the reward by comparing the model's predicted answer $\widehat{A}$ with the ground truth A.

The overall reward function is defined as:

$$R(O) = R_{\text{format}}(O) + R_{\text{acc}}(\widehat{A}; A). \quad (4)$$

RL Training. As mentioned in Sec. 3.1, we adopt the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) algorithm for RL on video QA tasks, using the proposed reward functions outlined in Eq. (4). Given an input video V and query Q, the model first generates G diverse candidate responses $\{O_1, \ldots, O_G\}$ with varied sampling configurations. Format and accuracy rewards (Eq. (4)) are then applied to each candidate response to compute their corresponding reward scores $\{R_1, \ldots, R_N\}$. Subsequently, the model (i.e., policy) π_θ is optimized using the GRPO objective as detailed in Eq. (2). This approach enables efficient RL training for video reasoning using only readily available video-question-answer triplets, with outcome-based rewards that are simple and fast to compute. Empirically, we find that our RL training design achieves comparable video reasoning performance using just $6K$ samples, compared to existing methods trained on large-scale SFT and RL datasets ($165K$ + $4K$), demonstrating the data and computational efficiency of VIDEO-RTS.

3.4 Dynamic Sparse-to-Dense Video Test-Time Scaling

With the RL training on video reasoning data, our model is able to generate a long chain-of-thought reasoning process to solve the video reasoning problem. However, as shown in Fig. 3, we find that after $6K$ training samples, adding many more video QA samples brings marginal improvements to the video reasoning performance. Inspired by the recent progress in test-time scaling technique (Snell et al., 2024; Wang et al., 2022; Yao et al., 2023) from the language community, we aim to save the excessive computational resources in the training stage and allocate them during the inference stage to improve the video reasoning capability. Given the redundant nature of video data (Wang et al., 2024c), an adaptive inference strategy with sparse-to-dense exploration can be both efficient and effective.

Algorithm 1: Sparse-to-Dense Video TTS

Input : Video $\mathbf{V}$ (N frames), query Q; policy π_θ; sample count m; initial frame budget n_{init}; maximum budget $n_{\max}$

Output : Predicted answer $\hat{A}$

$n \leftarrow n_{\text{init}}$;
while $n \leq n_{\max}$ **do**
  $\mathcal{V}(n) \leftarrow$ first n frames of $\mathbf{V}$;
  for $i \leftarrow 1$ **to** m **do**
    $O_i \leftarrow \pi_\theta(\mathcal{V}(n), Q)$;
    $\hat{A}_i \leftarrow \texttt{ExtractAnswer}(O_i)$;
  if $\hat{A}_1 = \hat{A}_2 = \cdots = \hat{A}_m$ **then**
    return $\hat{A} \leftarrow \hat{A}_1$;
  else if $n = n_{\max}$ **then**
    return $\hat{A} \leftarrow \texttt{MajorityVote}(\{\hat{A}_i\}_{i=1}^{m})$;
  $n \leftarrow \min(n * 2, n_{\max})$;

To this end, we propose a sparse-to-dense video test-time scaling strategy that iteratively refines the video reasoning process by scaling the frame inputs. Inspired by the majority voting method (Wang et al., 2022) in the NLP domain, we utilize the self-consistency of the MLLM as the signal of whether the model requires denser information for accurate video reasoning. Specifically, given the input video with n frames $V(n)$ and query Q, the RL-trained model π_θ generates m different responses $\{O_1, \ldots, O_m\}$ given m different sampling parameter for the MLLM. These responses include diverse reasoning processes on the video input and given query, which provide logical thinking from different angles under the current frame rate. Then, we extract the predicted answer $\{\widehat{A}_1, \ldots, \widehat{A}_m\}$ from each output and check whether different reasoning process leads to a unanimous answer prediction. If the diverse reasoning processes make a consensus, we consider the current temporal information sufficient, and we trust the current prediction. If the model generates conflicting predictions, we consider that the current temporal information is not enough for the model to generate an accurate response on the given video and query. Thus, we increase the frame rate and conduct the majority voting process iteratively until the model finds a consensus or it reaches the frame rate limit. We show the detailed algorithm of sparse-to-dense video test-time scaling in Algorithm 1. With the sparse-to-dense exploration during the inference stage, VIDEO-RTS adaptively allocates the computational budget for the sample with different temporal requirements and improves the video reasoning performance.

4 Experimental Setup

4.1 Evaluation Benchmarks

(1) **Video-Holmes** (Cheng et al., 2025) is a newly released and challenging benchmark designed to evaluate the complex video reasoning capabilities of MLLMs. It consists of 1837 questions sourced from 270 manually annotated suspense short films (ranging from 1 to 5 minutes), which span seven carefully curated tasks.

(2) **MMVU** (Zhao et al., 2025) is a comprehensive expert-level, multi-discipline benchmark for evaluating video understanding. We test on the val split of MMVU on multiple-choice QA format, which contains 625 QA samples that require expert-level reasoning on complex videos.

(3) **VideoMMMU** (Hu et al., 2025) is a multi-modal and multi-disciplinary video benchmark that evaluates LMMs' knowledge acquisition capability from videos. We use the standard split, which contains 900 video reasoning questions in perception, comprehension, and adaptation tasks.

(4) **Video-MME** (Fu et al., 2024b) is a recently proposed comprehensive evaluation benchmark for video analysis from short to long videos ((avg. 17 min)). We use the standard split of Video-MME, which contains 2700 expert-labeled QA pairs designed for both perception and reasoning tasks.

(5) **LongVideoBench (LVB)** (Wu et al., 2024) is a video QA benchmark that highlights referred reasoning questions, which are dependent on long frame inputs. We test on the public validation split, which contains 1337 video reasoning questions.

4.2 Evaluation Metrics

We evaluate VIDEO-RTS on all datasets under the multiple-choice QA setting. We utilize standard accuracy metrics for all experiments.

4.3 Training Data

We leverage CG-Bench (Chen et al., 2024a) as training data, which originally contains $12K$ MCQA data. Following Yu et al. (2025); Zheng et al. (2025), we filter out the samples that are too easy or too hard for effective learning. Specifically, we generate 8 responses per sample and calculate the difficulty based on the accuracy, samples with accuracy of either 0 or 1 are excluded. We finally sample a subset of $6K$ MCQA pair for training.

Method	#Frame	Video-Holmes	MMVU(mc)	Video-MMMU	LVB(val)	Video-MME
Proprietary MLLMs						
GPT-4o	-	42.0	75.4	61.2	66.7	71.9
Gemini 1.5 Pro	-	41.3	71.2	53.4	64.0	75.0
Open-Source General-Purpose MLLMs						
LLaVA-OV-7B	64	-	49.2	33.8	56.3	58.2
ViLA-1.5-8B	64	-	49.2	33.8	56.3	58.2
Qwen-2.5-VL-7B	≤ 768	27.8	59.2	47.4	56.0	**65.1**
Video Reasoning LLMs						
VideoMind-7B	> 64	–	–	–	–	58.2
VideoTree	64	–	54.2	47.8	52.3	56.1
Video-R1-7B	64	36.5	63.8	52.4	53.4	61.4
VIDEO-RTS-7B (ours)	**51.2**	**40.7**	**66.4**	**52.7**	**56.6**	63.0

Table 1: Comparison of the overall accuracy (%) with the state-of-the-art methods on five video reasoning tasks. We **highlight** the best performance model on 7B scale for each benchmark.

4.4 Implementation Details

We use Qwen-2.5-VL-7B (Bai et al., 2025a) as our base MLLM. To speed up the training process, we uniformly sample 32 frames for each video and resize the short side of the video to 224 resolution while keeping the original aspect ratio. For GRPO-related implementation, we reference the TRL (von Werra et al., 2020) library. We train our model with 1 epoch and only fine-tune the LLM's parameters. For 8 NVIDIA-H100 GPUs, the training takes approximately half a day to finish. For hyperparameters, we follow the recent works (Zhang et al., 2025a) and set the learning rate as $1e-6$ and the batch size as 16, the β for KL is set to 0.04. For all voting methods, we set the sample number m as 5. For evaluation, we set the base frame number as 32 as the default. For knowledge-focused benchmarks (MMVU, Video-MMMU), We set the max frame number as 64 for knowledge-focused benchmarks (MMVU, Video-MMMU) and 128 for general long video (reasoning) benchmarks (Video-Holmes, Video-MME, LVB).

5 Experimental Results

5.1 Comparison with State-of-the-Art

Tab. 1 shows a comparison of the existing works and VIDEO-RTS on five popular video reasoning benchmarks. We compare our methods with three types of models: leading proprietary MLLMs (OpenAI, 2024a; Team, 2024), open-source general-purpose MLLMs (Li et al., 2024; Bai et al., 2025a) and video reasoning LLMs (Liu et al., 2025; Wang et al., 2024c; Feng et al., 2025). Specifically, our approach achieves an accuracy of 40.7% on the Video-Holmes benchmark, outperforming the best open-source 7B model by a significant margin of 4.2%, and performing comparably to proprietary models such as Gemini 1.5 Pro and GPT-4o. These results validate the effectiveness of VIDEO-RTS towards complex video reasoning tasks. On the benchmarks that require expert-level reasoning ability over the complex videos (MMVU, Video-MMMU), VIDEO-RTS outperforms the SOTA video reasoning model trained on $169K$ total training samples (Video-R1 (Feng et al., 2025)) by 2.6% and 0.3% using only $6k$ samples. Meanwhile, our methods only utilize 42.8 and 45.2 average frames for inference, validating the frame efficiency of VIDEO-RTS that provides by the adaptive video test-time scaling strategy at inference stage. On the general video understanding benchmarks (LVB, Video-MME) that contains complex long video inputs, VIDEO-RTS significantly outperforms Video-R1 by 3.2 and 1.6 with less frame input (60.5 and 56.5) for video inference. Our method also outperforms the leading open-source general-purpose MLLM (Qwen-2.5-VL (Bai et al., 2025a)) on LVB with 92.2% average frames for evaluation. This result shows the efficiency and effectiveness of VIDEO-RTS on reasoning over complex video inputs. To sum up, our method achieves the best report performance among all open-source video reasoning methods

5.2 Analysis

Efficiency of the Pure-RL Training. In Tab. 2, we verify the efficiency and effectiveness of our pure-RL training with zero-shot CoT prompting on base MLLM (Bai et al., 2025a), large-scale SFT

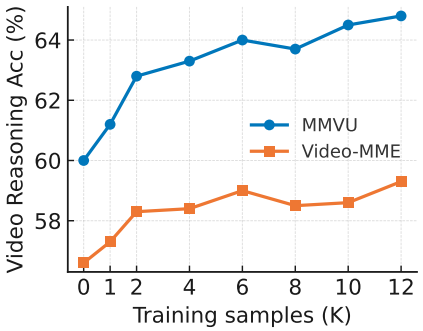

Figure 3: Analysis on the number of training samples for pure-RL training.

Reasoning	Data	MMVU	Video-MME
Zero-shot CoT	–	60.0	56.6
SFT	165K	63.5	55.4
SFT+RL	165K+4K	63.8	**59.3**
VIDEO-RTS (Ours)	6K	**64.0**	59.0

Table 2: Efficiency of our pure-RL training method.

Inference Method	MMVU	Video-MME
Pure-RL (vanilla)	64.0	59.0
+ Self-Consistency	64.6	60.6
+ Weighted Voting	64.2	60.4
+ Self-Reflection	64.4	59.3
+ S2D Video TTS (Ours)	**66.4**	**63.0**

Table 3: Comparison of sparse-to-dense video test-time scaling with different test-time scaling strategies. S2D refers to 'sparse-to-dense'.

Pure-RL	S2D Video TTS	MMVU	Video-MME
		60.0	56.6
✓		64.0	59.0
	✓	63.8	59.6
✓	✓	**66.4**	**63.0**

Table 4: Effectiveness of combining our pure-RL and adaptive video test-time scaling design for video reasoning tasks. The baseline model on the top row is zero-shot CoT on base MLLM.

and SFT+RL frameworks (Feng et al., 2025). With only 6K training data, our pure-RL method gets on par performance with the large-scale SFT+RL framework (trained on 165k+4K data) with only 3.6% samples, validating the training efficiency of VIDEO-RTS. What's more, our pure RL only requires the ground truth answer as a training signal, which skips the cumbersome data generation process for SFT on the video reasoning task. Meanwhile, in Fig. 3, we also analyze the gain in video reasoning performance with different numbers of training samples. We observe a sharp performance gain within the first $2K$ samples and continually get improvements on both MMVU and Video-MME until $6K$ samples. However, the pure-RL training seems to be saturated at $6K$ sample,s and scaling to more samples will degrade or get marginal performance. We also show the results with even more training samples in Tab. 6, which confirms this trend. We argue that the base MLLM possesses powerful reasoning capabilities derived from its pretraining and post-training stages. Our pure-RL training method helps the model become familiar with the video QA format and transfers the reasoning ability from language to complex video inputs. Thus, we propose saving the excessive training resources and allocating them to the inference stage to improve video reasoning capability while remaining efficient.

Effectiveness of Video-Specific TTS Design. In Tab. 3, we compare adaptive video test-time scaling with the popular test-time scaling (TTS) methods: self-consistency (Wang et al., 2022), weighted voting (Wan et al., 2025), self-reflection (Shinn et al., 2023). For weighted voting, we use the recent IXC-2.5-Reward (Zang et al., 2025) model as the verifier model. Results show that adaptive video test-time scaling outperforms the existing language-based TTS methods by 2.0% and 2.6% on MMVU and Video-MME benchmarks, respectively. This validates the adaptive design of adaptive video test-time scaling improves the reasoning capability for VIDEO-RTS on complex videos.

Vote Count Analysis of Video-TTS. In Tab. 5, we ablate the effectiveness of vote counts in S2D video TTS design. Moving from a single trajectory to five significantly improves the video reasoning performance on MMVU by 2.4% and Video-MME by 4.0%. This result indicates that five independent samplings already offer sufficient diversity for robust majority agreement while keeping VIDEO-RTS efficient. Pushing the pool to 10 or 20 adds even more low-probability reasoning chains, making the model's consensus less reliable and performance dips despite higher compute.

RL+TTS Yields Strong Video Reasoning Ability. In Tab. 4, we showcase the two main components of our method, pure-RL training and sparse-to-dense video test-time scaling, which are complementary for strong video reasoning capability. Individually, pure-RL training sharpens the reasoning ability of the base MLLM and brings significant gains

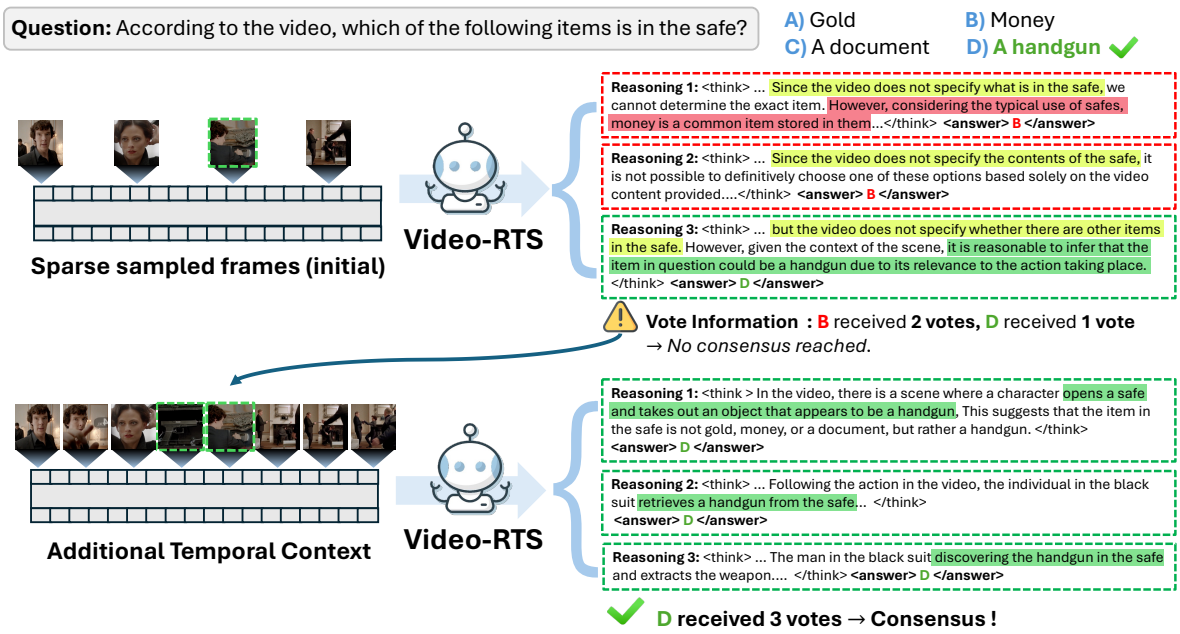

Figure 4: **Illustration of dynamic sparse-to-dense reasoning in VIDEO-RTS.** VIDEO-RTS identifies when the sampled visual information is insufficient for accurately reasoning about the input query (reasoning highlighted in yellow background), often leading to no consensus among intermediate reasoning steps and potentially inaccurate predictions (in red). VIDEO-RTS enables the model to adaptively refine its reasoning process (in green), through the proposed dynamic sparse-to-dense reasoning mechanism, achieving accurate and consensus-driven predictions.

#Votes	MMVU	Video-MME
1	64.0	59.0
5	66.4	63.0
10	66.2	63.1
20	65.8	62.6

Table 5: Ablation on vote count m in S2D Video-TTS. The vote 1 is VIDEO-RTS without test-time scaling.

compared to the zero-shot CoT prompting baseline. adaptive video test-time scaling alone also improves the performance, but its gains are limited by the reasoning capability of the base MLLM. When pure-RL training and S2D Video TTS are combined, improvements stack almost additively, pushing accuracy to 66.4% on MMVU and 63.0% on Video-MME.

Qualitative Analysis. In Fig. 4, we visualize qualitative results from VIDEO-RTS. Specifically, we show the effectiveness of the sparse-to-dense video inference process of VIDEO-RTS. In this example, given the query *"According to the video, which of the following items is in the safe"*, our model initially attempts to reason using a sparse frame set (i.e., a small number of sampled frames). As shown at the top, the reasoning (highlighted in yellow) lacks sufficient visual evidence, resulting in unclear and inconsistent predictions across multiple runs. In this case, VIDEO-RTS dynamically integrates additional frames into the inference process (bottom), allowing for more accurate and consistent reasoning by leveraging concrete visual cues (highlighted keyframes and reasoning steps are marked in green) to answer the visual query. This visualization showcases the adaptiveness of VIDEO-RTS during the inference stage and leads to accurate video reasoning.

6 Conclusion

We introduce VIDEO-RTS, the first work that systematically explores the combination of reinforcement learning and test-time scaling for video reasoning tasks. Instead of using long CoT data for SFT, we utilize a pure-RL training and propose an adaptive video inference strategy that allocates the excessive training compute to test time and improves the video reasoning capability. We validate the effectiveness of our model on four popular benchmarks, showing that VIDEO-RTS outperforms the recent video reasoning model by 2.4% in average accuracy with 3.6% training samples and fewer frames for inference. Importantly, we find that pure-RL training and adaptive video test-time scaling work synergistically, yielding the superior video-reasoning ability of VIDEO-RTS.

Acknowledgments

We thank the reviewers and area chair, as well as Justin Chen, David Wan, Ce Zhang and Yan-Bo Lin for their helpful discussions. This work was supported by DARPA ECOLE Program No. HR00112390060, NSF-AI Engage Institute DRL-2112635, ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, Capital One Research Award, the Accelerate Foundation Models Research program, Laboratory for Analytic Sciences via NC State University, and Sony Focused Research Award. The views contained in this article are those of the authors and not of the funding agency.

References

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, and 8 others. 2025a. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.

Shyamal Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the "Video" in Video-Language Understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. 2024a. Cg-bench: Clue-grounded question answering benchmark for long video understanding. Preprint, arXiv:2412.12075.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. Preprint, arXiv:2309.13007.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025. Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. 2025. Reinforcing video reasoning with focused thinking. Preprint, arXiv:2505.24718.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Preprint, arXiv:2501.12948.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. Preprint, arXiv:2501.03230.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024b. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. Preprint, arXiv:2405.21075.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. Preprint, arXiv:2501.13826.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749.

Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. 2023. Efficient movie scene detection using state-space transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18749–18758.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18198–18208.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. 2025. Bimba: Selective-scan compression for long-range video question answering. arXiv preprint arXiv:2503.09590.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. Journal of artificial intelligence research, 4:237–285.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision, pages 706–715.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2022. Revealing single frame bias for video-and-language learning. Preprint, arXiv:2206.03428.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326.

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

Yunxin Li, Xinyu Chen, Zitao Li, Zhenyu Liu, Longyue Wang, Wenhan Luo, Baotian Hu, and Min Zhang. 2025b. Veripo: Cultivating long reasoning in video-llms via verifier-gudied iterative policy optimization. arXiv preprint arXiv:2505.19000.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. 2025. Videomind: A chain-of-lora agent for long video reasoning. arXiv preprint arXiv:2503.13444.

Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, and Limin Wang. 2025. Videocap-r1: Enhancing mllms for video captioning via structured thinking. Preprint, arXiv:2506.01725.

Md Mohaiminul Islam and Gedas Bertasius. 2022. Long movie clip classification with state-space video models. European Conference on Computer Vision (ECCV).

OpenAI. 2024a. Gpt-4o system card. Preprint, arXiv:2410.21276.

OpenAI. 2024b. Openai o1 system card. Preprint, arXiv:2412.16720.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. Preprint, arXiv:2305.18290.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36:53728–53741.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. Preprint, arXiv:2303.11366.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun MA, and Chao Zhang. 2025. video-salmonn-o1: Reasoning-enhanced audio-visual large language model. arXiv preprint arXiv:2502.11775.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. Preprint, arXiv:2306.03314.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint, arXiv:2403.05530.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, and 1 others. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. arXiv preprint arXiv:2501.06186.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. Preprint, arXiv:2506.13654.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2025. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling. Preprint, arXiv:2408.17017.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. 2025a. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. Preprint, arXiv:2505.12434.

Xizi Wang, Feng Cheng, Ziyang Wang, Huiyu Wang, Md Mohaiminul Islam, Lorenzo Torresani, Mohit Bansal, Gedas Bertasius, and David Crandall. 2024a. Timerefine: Temporal grounding with time refining video llm. arXiv preprint arXiv:2412.09601.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. 2025b. Timezero: Temporal video grounding with reasoning-guided lvlm. arXiv preprint arXiv:2503.13377.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, and 1 others. 2024b. Internvideo2: Scaling video foundation models for multimodal video understanding. arXiv preprint arXiv:2403.15377.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, and 1 others. 2025c. Reinforcement learning for reasoning in large language models with one training example. arXiv preprint arXiv:2504.20571.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2024c. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. arXiv preprint arXiv:2405.19209.

Kangda Wei, Zhengyu Zhou, Bingqing Wang, Jun Araki, Lukas Lange, Ruihong Huang, and Zhe Feng. 2025. Premind: Multi-agent video understanding for advanced indexing of presentation-style videos. Preprint, arXiv:2503.00162.

Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1884–1894.

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

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 36:11809–11822.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. Preprint, arXiv:2503.14476.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, Kai Chen, Dahua Lin, and Jiaqi Wang. 2025. Internlm-xcomposer2.5-reward: A simple yet effective multi-modal reward model. arXiv preprint arXiv:2501.12368.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. A simple llm framework for long-range video question-answering. EMNLP.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025a. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025b. Tinyllava-video-r1: Towards smaller lmms for video reasoning. Preprint, arXiv:2504.09641.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Video instruction tuning with synthetic data. Preprint, arXiv:2410.02713.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. Preprint, arXiv:2501.12380.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. Preprint, arXiv:2505.14362.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv preprint arXiv:2406.04264.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. Preprint, arXiv:2504.10479.

Appendix

Our appendix consists of Limitations (Sec. A), Additional Implementation Details (Sec. B), Additional Quantitative Analysis (Sec. C), Detailed Prompts (Sec. D) and License and Artifact Usage (Sec. E).

A Limitations

Like other LLM-based video reasoning systems, our method may inherit societal or ethical biases from malicious content in the pretraining data of the base (M)LLMs. However, since VIDEO-RTS mitigates this risk through a consensus-based adaptive inference strategy, where biased outputs under certain sampling conditions can be counteracted by more neutral ones, enhancing fairness. Additionally, our method relies on relatively small-scale video reasoning data, making it more practical to filter harmful samples compared to large-scale SFT and RL setups. In future work, we aim to further investigate data quality and develop fairness-aware video reasoning models.

B Additional Implementation Details

For training data, we sample from the CG-Bench dataset (Chen et al., 2024a), which is a All results are under same, fixed random seed with single runs. For the majority voting in all test-time scaling method, we sample with temperature ($\tau_i = 0.7 + 0.1\,i$) and nucleus threshold ($p_i = \max(0.5,\ 0.9{-}0.1\,i)$). Generation is capped at 1024 tokens and stops early at the sentinel token `</answer>`. During evaluation, following Video-R1 (Feng et al., 2025), we set the max frame resolution to $256 * 28 * 28$.

C Additional Quantitative Results

C.1 Training Data Analysis

We use CG-Bench as the primary training source, which contains $1.2k$ videos with $12k$ multiple-choice question-answering (MCQA) samples. Following Yu et al. (2025); Zheng et al. (2025), we filter out the samples that are either too easy or too difficult to support more effective learning. Specifically, we generate 8 responses per sample and calculate their difficulty based on accuracy; samples with accuracy of 0 or 1 are excluded. To ensure a diverse set of video types, we then randomly sample an equal number of examples from each video, resulting in a final training set of $6K$ samples. In Tab. 6, we report the impact of different training data selection strategies and show that the manually annotated CG-Bench data (Chen et al., 2024a) outperforms the larger LLaVA-Video-178K data (Zhang et al., 2024c). Meanwhile, filtering out overly easy or hard samples benefits RL training, further improves performance, and speeds up the training process.

Data Source	# Sample	Filter	MMVU	V-MME
Zero-shot	–		60.0	56.6
LLaVA-V	35k		63.5	55.4
CG-Bench	12k		63.2	**58.3**
LLaVA-V	6k	✓	63.8	**59.3**
CG-Bench	6k	✓	64.0	**59.0**

Table 6: Training data analysis.

C.2 Evidence for Performance Saturation with Increasing Training Samples

Fig. 3 already suggests a saturation point after roughly $6K$ training examples. To further investigate this trend, we extend training to $12K$ samples by combining the original CG-BENCH with carefully deduplicated examples from the public LLAVA-VIDEO-178K corpus (Zhang et al., 2024c). Sec. C.2 shows that beyond 6 K examples, accuracy gains plateau (and occasionally regress), highlighting the importance of data-efficient methods such as ours.

# Training Sample	15 K	20 K	25 K
MMVU	64.1	63.8	63.5
Video-MME	58.8	59.1	59.2

Table 7: RL training with more samples.

C.3 Random Seeds Analysis

To verify that our results are not sensitive to the particular subset chosen (random sample $6K$ samples from CG-Bench), we repeat the sampling procedure with five distinct random seeds. After the RL fine-tuning step, the MMVU accuracy averages 64.2 ± 0.3 across the five runs, closely matching

the 64.0 reported in Tab. 4. This confirms that the observed performance is robust to the exact choice of training samples.

C.4 Sparse-to-Dense Video TTS on Other MLLM Backbone

To further validate the effectiveness of our framework, we conduct additional experiments using InternVL-3(Zhu et al., 2025), which is also a leading MLLM that has already been trained on reasoning data using RL. We directly evaluate our S2D Video TTS method using the InternVL-3-8B model on a video reasoning benchmark and demonstrate that our S2D Video TTS can achieve a 1.5% improvement in MMVU accuracy, validating the generalization ability of Video-RTS.

D Detailed Prompts

We provide detailed prompts for VIDEO-RTS in Tab. 8. We use the same format of prompts for training and evaluation.

E License and Artifact Usage

E.1 License

We use standard licenses from the community and provide the following links to the licenses for the datasets, codes, and models that we used in this paper:

TRL: Apache

Video-R1: Video-R1

Open-R1-Video: Apache

Qwen-2.5-VL: Apache

CG-Bench: CG-Bench

Video-MMMU: MIT

Video-MME: Video-MME

MMVU: MMVU

LongVideoBench: CC-BY-NC-SA 4.0

E.2 Artifact Usage

The use of existing artifacts is consistent with their intended use in We will make our code and models publicly accessible and all created artifacts will be only for research purposes and should not be used outside of research contexts.

Table 8: **VIDEO-RTS detailed prompt.**

User
A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>.
Video: `Video Tokens`
Question: `Question`
Options: A: `Option-A`. B: `Option-B`. C: `Option-C`. D: `Option-D`.....
Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.')