

---

# Video-R1: Reinforcing Video Reasoning in MLLMs

---

Kaituo Feng<sup>1</sup>, Kaixiong Gong<sup>1</sup>, Bohao Li<sup>2</sup>, Zonghao Guo<sup>3</sup>, Yibing Wang<sup>4</sup>,  
Tianshuo Peng<sup>1</sup>, Benyou Wang<sup>2</sup>, Xiangyu Yue<sup>1\*</sup>

<sup>1</sup>CUHK MMLab, <sup>2</sup>CUHK (SZ), <sup>3</sup>Tsinghua University, <sup>4</sup>UCAS

<https://github.com/tulerfeng/Video-R1>

## Abstract

Inspired by DeepSeek-R1’s success in eliciting reasoning abilities through rule-based reinforcement learning (RL), we introduce Video-R1 as the first attempt to systematically explore the R1 paradigm for eliciting video reasoning within multi-modal large language models (MLLMs). However, directly applying RL training with the GRPO algorithm to video reasoning presents two primary challenges: (i) a lack of temporal modeling for video reasoning, and (ii) the scarcity of high-quality video-reasoning data. To address these issues, we first propose the T-GRPO algorithm, which encourages models to utilize temporal information in videos for reasoning. Additionally, instead of relying solely on video data, we incorporate high-quality image-reasoning data into the training process. We have constructed two datasets: Video-R1-COT-165k for SFT cold start and Video-R1-260k for RL training, both comprising image and video data. Experimental results demonstrate that Video-R1 achieves significant improvements on video reasoning benchmarks such as VideoMMU and VSI-Bench, as well as on general video benchmarks including MVBench and TempCompass, etc. Notably, Video-R1-7B attains a 35.8% accuracy on video spatial reasoning benchmark VSI-bench, surpassing the commercial proprietary model GPT-4o. All codes, models, data are released.

## 1 Introduction

Recent advancements in rule-based Reinforcement Learning (RL) [11] have significantly enhanced the reasoning capabilities of Large Language Models (LLMs) [10, 6]. In particular, DeepSeek-R1 [6] has demonstrated that carefully designed RL pipelines can lead to emergent and robust reasoning abilities with long chain-of-thoughts (COT) in text-based domains. Motivated by this success, several recent efforts have explored extending RL training to Multimodal Large Language Models (MLLMs). Notable examples include Kimi k1.5 [21] and Skywork R1V [25], which apply RL to improve reasoning over image-text pairs. However, despite these early explorations, the domain of video reasoning in MLLMs remains largely unexplored.

To bridge this gap, we present Video-R1 model, as the first attempt to systematically investigate eliciting strong video reasoning based on the R1 paradigm. However, directly applying RL training with the Group Relative Policy Optimization (GRPO) algorithm [19] to video reasoning introduces two fundamental challenges: First, original GRPO lacks explicit reward signals for encouraging temporal reasoning in a video. Without explicit temporal awareness, the model may take shortcuts for reasoning, focusing on on a single frame or snapshot rather than reasoning over time (see Figure 1 for example, one previous work Video-UTR [26] also reveals a similar issue). The absence of temporal inductive bias can cause the learned reasoning strategies to "shortcut" the process—relying on superficial visual patterns, rather than engaging in deeper and temporally grounded reasoning. This could ultimately hinder generalization to more complex or diverse video reasoning tasks.

---

\*Corresponding Authors

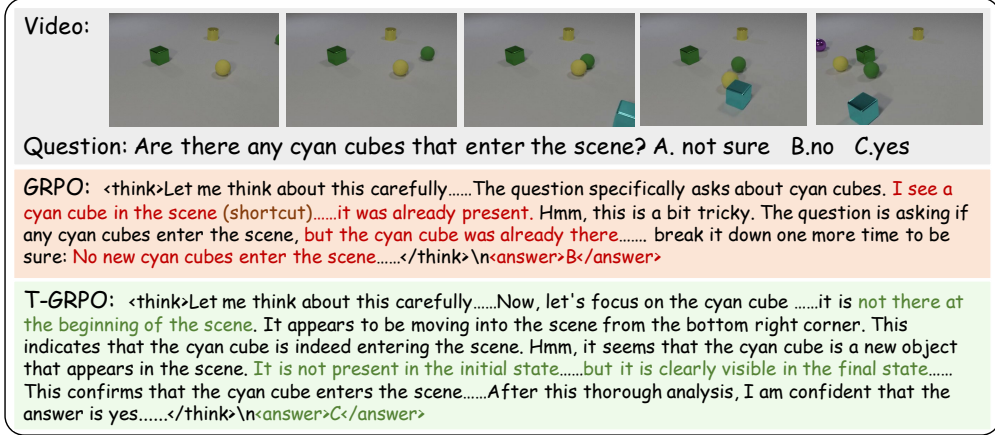


Figure 1: Reasoning paths of Video-R1 trained by GRPO and our proposed T-GRPO on test samples. Without explicit temporal modeling, models may learn sub-optimal video reasoning patterns by taking shortcuts, therefore failing to generalize well.

The second issue lies in the scarcity of high-quality video reasoning training data, especially problems that demand strong reasoning ability or involve long reasoning path. Most existing video training datasets mainly focus on simple recognition tasks, rather than reasoning. This scarcity makes it difficult to expose the model to diverse, challenging reasoning patterns during training, limiting the effectiveness of RL and hindering the emergence of robust reasoning behaviors.

To address these challenges, we propose two key solutions. First, we propose T-GRPO, an extension of the original GRPO algorithm that explicitly encourages temporal reasoning. During training, the model is presented with both temporally ordered and randomly shuffled frame sequences, producing two sets of responses. A positive reward is assigned only when the proportion of correct answers from the ordered sequence exceeds that from the shuffled one. This strategy encourages the model to exploit temporal reasoning policy rather than relying on shortcuts derived from isolated frames.

Besides, to tackle the scarcity of high-quality video reasoning data, we strategically introduce image-based reasoning data as part of training data. We construct two datasets: Video-R1-COT-165k for SFT cold start and Video-R1-260k for RL training. The image data serves as a valuable foundation for training general reasoning skills, while the curated video samples provide the temporal complexity needed for video understanding. This hybrid training setup not only alleviates the data bottleneck but also enables the model to transfer reasoning skills learned from static images to dynamic video contexts. Combined with T-GRPO, this approach equips Video-R1 with stronger, more generalizable video reasoning capabilities.

Our experiments show that Video-R1 achieves consistent and significant improvements across a suite of challenging video reasoning benchmarks, including VSI-Bench [24], VideoMMMU [8], MMVU [32], MVBench [13], TempCompass [18], and VideoMME [4]. Notably, Video-R1-7B attains 35.8% accuracy on VSI-Bench, a challenging video spatial reasoning benchmark, outperforming even proprietary models like GPT-4o [9]. These results suggest that with carefully designed algorithms and data pipelines, RL can indeed unlock complex temporal reasoning capabilities in MLLMs, similar to the breakthroughs seen in the text domain. Our contributions can be summarized as follows:

- We propose **Video-R1**, as the first attempt to systematically to explore developing video reasoning MLLMs based on the R1 paradigm. To support training, we construct two reasoning datasets: **Video-R1-COT-165k** for SFT and **Video-R1-260k** for RL training, incorporating both image and video reasoning samples. We hope that Video-R1 will serve as a foundation for future research on video reasoning.
- To address the lack of temporal modeling in existing RL methods, we introduce **T-GRPO**, a novel training algorithm that encourages the model to utilize temporal information by contrasting reasoning performance over ordered and shuffled video frames.
- Extensive experiments on multiple video benchmarks, such as Video-MMMU, VSI-Bench, MVBench, etc, demonstrate the effectiveness of our approach. Notably, **Video-R1-7B** achieves **35.8%** accuracy on VSI-Bench, outperforming the proprietary GPT-4o model.

## 2 Related Works

### 2.1 Multimodal Large Language Models for Video

Video understanding is an essential capability for Multimodal Large Language Models (MLLMs), enabling them to interpret and reason over dynamic visual content [20, 2, 23, 29, 30]. In recent years, a number of MLLMs have been developed specifically to advance progress in video understanding tasks. For example, LLaMA-VID [15] proposes a dual-token strategy (context and content tokens) to compress video input representations, enabling vision-language models to efficiently handle long videos while retaining essential visual information. VideoLLaMA2 [2] enhances video-language modeling by introducing spatial-temporal convolution for better dynamic understanding and an audio branch to integrate multimodal cues for richer video comprehension. LongVA [28] extends the context window of language backbones to process significantly longer video sequences without specialized video training, offering a language-centric solution to long-range temporal reasoning. VISA [23] introduces a knowledge-driven video object segmentation task that combines world knowledge with object tracking, addressing implicit, complex video queries through a segmentation-enabled multimodal LLM. These advancements highlight the potential of MLLMs in advancing video understanding. However, most prior works have primarily focused on video perception tasks. The development of MLLMs with strong video reasoning capabilities remains largely unexplored.

### 2.2 Large Language Model Reasoning

The reasoning abilities of Large Language Models (LLMs) have been a focal point of recent research, aiming to enhance their capacity to perform complex, multi-step problem-solving tasks [22, 31, 27]. Unlike earlier approaches that rely on dense, step-level supervision or learned reward models to supervise reasoning paths [5, 14], DeepSeek-R1 initiated a new wave of interest in rule-based reinforcement learning, demonstrating that even coarse, outcome-only rewards can effectively elicit strong reasoning behavior [6]. Its success showed that with a carefully designed reward structure and policy optimization strategy, models could learn to generate long COT without requiring intermediate supervision. Following this paradigm, several recent efforts have attempted to reproduce R1’s success. For example, Open Reasoner Zero [7] and Kimi k1.5 [21] explore similar rule-based RL pipelines to enhance reasoning in the text and image domains, respectively. However, despite encouraging progress, few prior work has explored how to extend this approach to the video domain. Bridging this gap remains an open challenge and a promising direction for expanding the boundaries of reasoning models.

## 3 Methods

### 3.1 Dataset Construction

High-quality training data plays a crucial role in reinforcing video reasoning capabilities in MLLMs. In this section, we will introduce how we curate Video-R1-260k for RL training and Video-R1-COT-165k for SFT cold start.

**Data Collection and Curation.** To overcome the scarcity of high-quality video reasoning training data, we strategically introduce image-based reasoning data as part of training data. The image-based data serves primarily to teach the model a broad range of reasoning skills, covering various difficulty levels and domains such as math, spatial logic, expert-level knowledge, etc. These samples help the model develop generalized reasoning abilities in static contexts. In contrast, the video-based data is primarily used to train the model’s ability to perform temporal reasoning—including understanding event progression, capturing frame-to-frame dependencies, and drawing inferences based on motion and causal dynamics over time.

We collect data from a variety of public datasets and carefully sample and balance the proportion of each subset. The final composition of the Video-R1-260k dataset is illustrated in Figure 2. The distribution of Video-R1-260k dataset can be roughly categorized as follows:

- **General (Video, 116k):** A diverse set of open-domain video data, covering everyday scenarios and designed to build temporal comprehension and reasoning abilities.

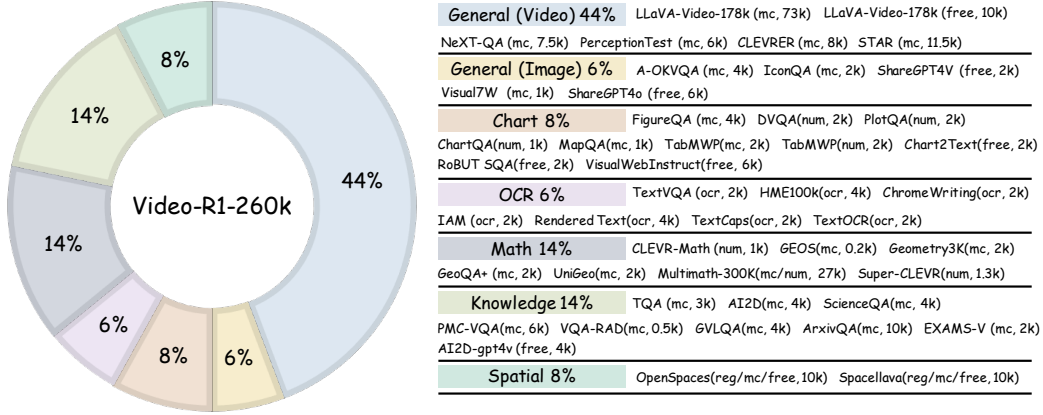


Figure 2: The data distribution of our Video-R1-260k dataset.

- **General (Image, 15k):** A general-purpose image question-answering data, used to provide basic visual understanding.
- **Chart (Image, 21k):** Visual reasoning over charts, line graphs and scientific figures, focusing on data interpretation and quantitative logic.
- **OCR (Image, 16k):** Facilitate reasoning tasks that require recognizing and interpreting embedded textual content such as signs, forms, or documents.
- **Math (Image, 37k):** Image-based math reasoning questions, including formulas, geometry diagrams, and multi-step symbolic reasoning.
- **Knowledge (Image, 37k):** Visual commonsense and Multi-discipline reasoning tasks, testing the model’s ability to integrate world knowledge with visual cues.
- **Spatial (Image, 20k):** Tasks that require understand spatial information for reasoning.

**COT Annotation.** To facilitate an effective SFT cold start, we leverage Qwen2.5-VL-72B [1] to generate COT rationales for the samples in Video-R1-260k. After applying basic rule-based filtering to remove low-quality or inconsistent outputs, we obtain a high-quality CoT dataset, Video-R1-COT-165k, which is used for the cold-start SFT stage.

**Data Type and Rule-based Reward Design.** Since our reinforcement learning framework follows the rule-based reward paradigm of DeepSeek-R1, it is crucial to ensure that the reward signals are both reliable and precise. To this end, the majority of our training data is designed around tasks with clearly verifiable outputs, such as multiple-choice and numerical answer formats. This allows for accurate reward computation using simple rules, ensuring stable and effective RL training.

However, to increase the model’s flexibility and its ability to generalize across diverse tasks and formats, we also incorporate a smaller portion of other data types. These include free-form generation, OCR tasks, and regression problems, which are essential for adapting to real-world applications and broader datasets.

The data types and corresponding reward functions are summarized as follows:

- **Multiple Choice:** The reward is assigned based on whether the predicted option matches the ground-truth answer.
- **Numerical QA:** A binary reward is given depending on whether the predicted number exactly matches the reference value.
- **OCR:** We compute the reward using the Word Error Rate (WER), measuring the edit distance between the predicted and reference text.
- **Free-form QA:** The reward is calculated as the average of ROUGE-1, ROUGE-2, and ROUGE-L scores between the model’s output and the ground-truth answer.

- **Regression:** The closer the predicted value is to the ground truth, the higher the reward, calculated as one minus their relative error.

### 3.2 Temporal Group Relative Policy Optimization (T-GRPO)

While GRPO has proven effective in text-based reasoning, it lacks explicit reward signals for temporal reasoning—making it insufficient for training MLLMs to reason over videos. To address this, we propose Temporal Group Relative Policy Optimization (T-GRPO), which introduces a contrastive reward mechanism that explicitly encourages temporal reasoning.

The core idea behind T-GRPO is to compare the model’s performance on the same video question when frames are provided in two different orders: (1) the *temporally ordered* sequence, and (2) a *randomly shuffled* version. For each input question, we generate two groups of responses  $\{o_i\}_{i=1}^G$  and  $\{\tilde{o}_i\}_{i=1}^{\tilde{G}}$  using the ordered and shuffled frame inputs, respectively.

Let  $p$  and  $\tilde{p}$  denote the proportion of correct answers in each group. We then define a temporal reward coefficient  $r_t$  as:

$$r_t = \begin{cases} \alpha, & \text{if } p > \mu \cdot \tilde{p} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where  $\alpha$  and  $\mu$  are hyper-parameters. Here we set  $\alpha = 0.3$  and  $\mu = 0.8$ .

This contrastive design encourages the model to perform better when the video is presented in correct temporal order than when it is shuffled. The model is only granted this positive reward if its current reasoning strategy for a given question demonstrates a reliance on temporal information.

Importantly,  $r_t$  is only applied to correct responses to ensure meaningful positive advantages. Applying it to all responses would dilute the reward signal and hinder effective learning. In other words, when the model’s reasoning policy successfully relies on temporal patterns, correct responses are reinforced with a higher reward, while incorrect ones remain unaffected.

Formally, the temporal-augmented reward is defined as:

$$r_i^{\text{T-GRPO}} = \begin{cases} r_i + r_t, & \text{if } o_i \text{ is correct} \\ r_i, & \text{otherwise} \end{cases} \quad (2)$$

This reward shaping ensures that when the model answers correctly under a temporal setting but fails to outperform the shuffled baseline, it receives no additional reward—pushing the optimization toward adopting a more temporally aware reasoning policy.

The advantage  $A_i$  is computed over the updated rewards within each group, as in DeepSeek R1:

$$A_i = \frac{r_i^{\text{T-GRPO}} - \text{mean}(\{r_j^{\text{T-GRPO}}\})}{\text{std}(\{r_j^{\text{T-GRPO}}\})} \quad (3)$$

The final policy update follows the clipped surrogate objective of GRPO:

$$\mathcal{J}_{\text{T-GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\}} \left[ \frac{1}{G} \sum_{i=1}^G \left( \min \left( \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \quad \text{clip} \left( \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right) \right] \quad (4)$$

By explicitly comparing the model’s performance under ordered and shuffled inputs, T-GRPO introduces a contrastive training signal that drives the model to prefer reasoning strategies that leverage temporal patterns. It is worth noting that T-GRPO is only employed for video-based inputs in the training process of Video-R1.

### 3.3 Training Strategies

We adopt Qwen2.5-VL-7B [1] as the base MLLMs for training. Similar to DeepSeek R1, The training process is conducted in two stages: SFT cold start followed by RL training. For these two stages, we both adopt image-video mixed training strategy.

In the first stage, we perform SFT on the Video-R1-COT-165k dataset, which contains chain-of-thought (CoT) annotated samples derived from both image and video inputs. This step serves as a cold-start initialization, equipping the model with basic reasoning capabilities across a variety of modalities. The resulting model is denoted as Qwen2.5-VL-7B-SFT.

In the second stage, we further train the Qwen2.5-VL-7B-SFT model on the broader Video-R1-260k dataset using our proposed T-GRPO algorithm. This reinforcement learning phase is designed to guide the model beyond the rigid, pattern-matching behavior induced by supervised fine-tuning, encouraging it to freely explore and internalize more effective, generalizable reasoning strategies. The resulting model is denoted as Video-R1-7B.

To further enhance the quality of reasoning, we introduce a length-based reward to regulate the length of the model’s output. Specifically, this mechanism aims to strike a balance between encouraging deeper reasoning and preventing overthinking. For each reasoning path  $o_i$ , if the predicted answer is correct and the response length falls within a predefined interval  $[l_{\min}, l_{\max}]$ , the model receives an additional reward  $r_l = \omega$ . Formally:

$$r_i = \begin{cases} r_i + \omega, & \text{if } o_i \text{ is correct and } l_{\min} \leq \text{len}(o_i) \leq l_{\max} \\ r_i, & \text{otherwise} \end{cases} \quad (5)$$

This reward encourages the model to think deeply without overthinking. In this paper, we set  $\omega = 0.1$ ,  $l_{\min} = 320$  and  $l_{\max} = 512$ .

### 3.4 Aha Moment in Video Reasoning

One of the most intriguing outcomes of reinforcement learning in Video-R1 is the emergence of self-reflection reasoning behaviors, commonly referred to as “aha moments”. These moments occur when the model departs from a straightforward answer path, pauses to reassess its prior steps, and ultimately converges on a more accurate or logically sound solution. In Video-R1, we observe that the model occasionally revisits its interpretation of the video, especially when faced with ambiguous temporal cues or multi-step inference. Instead of committing to its initial plan, the model may generate reasoning traces that reflect uncertainty, reconsideration, or alternate strategies. For example, in some cases, the model begins with a plausible assumption, identifies an inconsistency mid-way, and then issues a correction by re-analyzing earlier frames or re-evaluating temporal relationships. These behaviors suggest that the model is not merely executing memorized patterns, but is actively engaging in internal feedback loops—re-examining evidence and adjusting its conclusions accordingly. Some examples are demonstrated in Figure 3, 4.

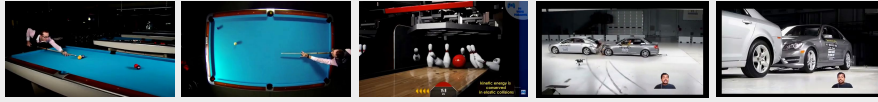
## 4 Experiments

### 4.1 Setup

**Benchmarks.** We evaluate our model on six video benchmarks: VSI-Bench [24], VideoMMMU [8], MMVU [32], MVBench [13], TempCompass [18], and VideoMME [4]. Among them, the first three are video reasoning benchmarks, which focus primarily on assessing the model’s reasoning capabilities in video understanding. Specifically, VSI-Bench evaluates spatial reasoning over videos, while VideoMMMU and MMVU assess knowledge-intensive video question answering that requires integrating visual and knowledge information across time. For MMVU, we evaluate on its multiple-choice question set for stability and consistency.

The latter three, MVBench, TempCompass, and VideoMME, are general-purpose video understanding benchmarks, which include a mixture of perception and reasoning tasks. Although not exclusively focused on reasoning, they provide a broader assessment of the model’s overall video comprehension abilities, and include subsets that require temporal or logical reasoning.

Video:



Question: Which move motion in the video lose the system energy?  
A. Third one B. Second one C. All D. First one E. Two of the three

Video-R1-7B: <think>Let me think about this carefully. The question is asking which move loses the system energy. To answer this, I need to analyze each of the three moves shown in the video and determine if any of them result in a loss of system energy.

First, let's break down the first move. In the first move, a person is playing pool. The cue ball strikes the other balls, and they scatter across the table. This is a collision where kinetic energy is conserved. The balls move, but the total kinetic energy of the system remains the same. So, this move does not lose system energy.

Next, let's consider the second move. In the second move, a person is bowling. The bowling ball strikes the pins, and the pins scatter. This is also a collision where kinetic energy is conserved. The pins move, but the total kinetic energy of the system remains the same. So, this move does not lose system energy either.

Finally, let's look at the third move. In the third move, two cars are shown in a crash test. The cars collide, and the energy is dissipated in the form of deformation and heat. This is a collision where kinetic energy is not conserved. The cars lose energy in the form of deformation and heat, which is a loss of system energy. (Perception and Reasoning)

Hmm, it seems that the third move is the only one that results in a loss of system energy. The first and second moves conserve kinetic energy, while the third move dissipates energy.

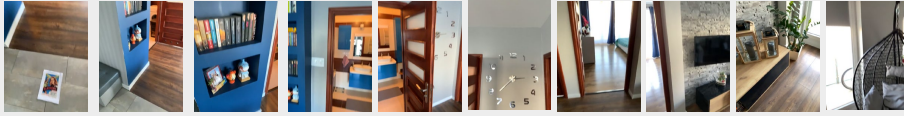
Oh, I see. The correct answer is the third move, which loses the system energy.

Let's verify this one last time. The first and second moves conserve energy, while the third move dissipates energy. This confirms that the third move is the correct answer. Therefore, the answer is A. Let's break it down one last time to ensure. The first and second moves conserve energy, while the third move dissipates energy. This is a clear indication that the third move is the correct answer. </think> (Reflection)

<answer>A</answer>

Figure 3: An example of Video-R1-7B's reasoning output on MMVU Benchmark.

Video:



Question: You are a robot beginning at the blue built-in bookshelf facing the blue built-in bookshelf. You want to navigate to the bathroom sink. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward until the door of the bathroom. 3. [please fill in] 4. Go forward until the sink. You have reached the final destination.  
A. Turn Back, Turn Right B. Turn Left, Turn Left C. Turn Right, Turn Left D. Turn Back, Turn Left

Video-R1-7B: <think>Let me think about this carefully. The goal is to navigate from the blue built-in bookshelf to the bathroom sink. The first step is to determine the direction from the bookshelf to the bathroom.

1. The blue built-in bookshelf is on the left side of the room. The bathroom is on the right side of the room. So, I need to turn right from the bookshelf to head towards the bathroom.

2. After turning right, I need to go forward until I reach the door of the bathroom. This means I need to continue moving forward in the direction I just turned.

3. Once I reach the bathroom door, I need to turn left to enter the bathroom. This is because the sink is typically located on the left side of the bathroom.

So, the sequence of actions should be: 1. Turn right from the bookshelf. 2. Go forward until the bathroom door. 3. Turn left to enter the bathroom. 4. Go forward until the sink. (Perception and Reasoning)

Hmm, this seems to align with the sequence of actions in option C. Let's break it down one more time to verify.

Option A: Turn Back, Turn Right - This would mean going back to the bookshelf and then turning right, which doesn't make sense for reaching the bathroom.

Option B: Turn Left, Turn Left - This would mean turning left from the bookshelf and then turning left again, which also doesn't make sense.

Option D: Turn Back, Turn Left - This would mean going back to the bookshelf and then turning left, which is not the correct path. (Reflection)

Oh, I see. The correct sequence is indeed in option C. Let's confirm one last time. The correct sequence is to turn right, go forward to the bathroom door, turn left, and go forward to the sink. This matches option C. Final answer: C  
Let's verify one last time. Yes, C is the correct answer. </think>

<answer>C</answer>

Figure 4: An example of Video-R1-7B's reasoning output on VSI-Bench.

Table 1: Performance of different models on benchmarks.

Models	Frames	Video Reasoning Benchmark			Video General Benchmark		
		VSI-Bench	VideoMMMU	MMVU (mc)	MVBench	TempCompass	VideoMME (wo sub)
GPT-4o [9]	-	34.0	61.2	75.4	-	-	71.9
LLaMA-VID [15]	-	-	-	-	41.9	45.6	-
VideoLLaMA2 [2]	-	-	-	44.8	54.6	-	47.9
LongVA-7B [28]	-	29.2	23.9	-	-	56.9	52.6
VILA-1.5-8B [16]	-	28.9	20.8	-	-	58.8	-
VILA-1.5-40B [16]	-	31.2	34.0	-	-	-	<b>60.1</b>
Video-UTR-7B [26]	-	-	-	-	58.8	59.7	52.6
LLaVA-OneVision-7B [12]	-	32.4	33.8	49.2	56.7	-	58.2
Kangaroo-8B [17]	-	-	-	-	61.1	62.5	56.0
Qwen2.5-VL-7B (COT)	16	27.7	47.8	59.2	57.4	72.2	53.1
Qwen2.5-VL-7B-SFT	16	31.8	47.4	61.3	59.4	69.2	52.8
Qwen2.5-VL-7B (COT)	32	30.1	48.1	60.0	59.0	72.6	56.6
Qwen2.5-VL-7B-SFT	32	33.3	49.4	63.5	60.5	69.9	55.4
Video-R1-7B	16	34.6	49.8	<b>64.2</b>	62.7	72.6	57.4
Video-R1-7B	32	<b>35.8</b>	<b>52.3</b>	63.8	<b>63.9</b>	<b>73.2</b>	59.3

For all evaluations, we follow the decoding configuration used in the official Qwen2.5-VL demo, with top\_p = 0.001 and temperature = 0.01.

**Training Details.** We train our model using 4 NVIDIA H20 (96GB) GPUs. For efficiency considerations, we limit the maximum number of video frames to 16 during training. Each frame is processed at a resolution of  $128 \times 28 \times 28$ . During inference, we increase the frame resolution to  $256 \times 28 \times 28$  and frames to 16~32 to enhance performance.

We first perform supervised fine-tuning on the Video-R1-COT-165k dataset for one epoch to obtain the Qwen2.5-VL-7B-SFT model. This is followed by RL training on the Video-R1-260k dataset to produce the final Video-R1 model. Due to current computational resource limitations, we train the model for only 1k RL steps.

Surprisingly, even within this limited training budget, the model exhibits significant improvements in video reasoning performance, indicating the strong effectiveness of both our data design and algorithm. In future work, we plan to scale up training with more frames and training steps to further boost performance.

## 4.2 Main Results

As shown in Table 4.1. Our experimental results across six benchmarks validate the effectiveness of Video-R1 in video reasoning and general video understanding tasks. The key findings are summarized as follows.

**Superior Performance of Video-R1.** Video-R1 significantly outperforms previous models across most benchmarks, with particularly strong gains on video reasoning tasks such as VSI-Bench, VideoMMMU, and MMVU. Notably, on VSI-Bench, which focuses on spatial reasoning in videos, Video-R1-7B achieves a new state-of-the-art accuracy of 35.8%, surpassing GPT-4o, a proprietary model, while using only 32 frames and 7B parameters. This highlights the necessity of explicit reasoning capability in solving video tasks, and confirms the effectiveness of reinforcement learning for video tasks.

**SFT Memorizes, RL Generalizes.** We observe that the SFT model Qwen2.5-VL-7B-SFT does not consistently improve performance across benchmarks. In some cases, such as VideoMME



Table 2: Ablation Study.

Models	Frames	Video Reasoning Benchmark			Video General Benchmark		
		VSI-Bench	VideoMMMU	MMVU (mc)	MVBench	TempCompass	VideoMME (wo sub)
Video-R1-7B-wo-image	16	32.3	45.8	60.6	60.9	69.8	53.8
Video-R1-7B-wo-temporal	16	32.7	48.3	62.1	61.1	71.3	54.5
Video-R1-7B	16	34.6	49.8	64.2	62.7	72.6	57.4

and TempCompass, performance even slightly drops after SFT, likely due to overfitting or limited generalization in unseen scenarios [3]. In contrast, after only 1k steps of reinforcement learning, Video-R1 achieves significant performance boosts across all benchmarks, especially in reasoning-heavy settings. This clearly demonstrates the strength of our RL framework and emphasizes the importance of reinforcement learning in unlocking generalizable video reasoning capability.

**More Frames Lead to Better Reasoning.** When increasing the number of input frames from 16 to 32, we observe performance improvements on almost all benchmarks. This indicates that both longer context and richer temporal information contribute positively to the model’s reasoning performance. The improvements are particularly notable in benchmarks like VideoMMMU and VideoMME, where longer video sequences offer more clues for inference. These results suggest that developing models capable of understanding and reasoning over longer video inputs is a promising and necessary direction for future research.

### 4.3 Ablation Study

To better understand the contributions of individual components in our method, we conduct an ablation study by evaluating two variants of our model: Video-R1-7B-wo-image, which removes all image-based data during training and relies solely on video data, and Video-R1-7B-wo-temporal, which replaces our proposed T-GRPO algorithm with the original GRPO method that lacks explicit temporal reasoning modeling.

As shown in Table 2, both ablated models perform worse than the full Video-R1-7B across all benchmarks. In particular, removing image data leads to a noticeable drop in performance on both video reasoning and general benchmarks, indicating that image-based samples play a crucial role in bootstrapping general reasoning ability and providing diverse supervision signals. Similarly, without temporal-aware training via T-GRPO, the model struggles to fully leverage temporal cues, resulting in weaker performance on benchmarks. These ablations validate the effectiveness of our proposed methods.

### 4.4 Training Curves

Figure 5 illustrates the RL training dynamics of Video-R1. We track three key indicators throughout the training process: accuracy reward, temporal reward  $r_t$ , and response length.

As shown in Figure 5(a), the accuracy reward exhibits a generally upward trend, indicating that the model continuously improves its ability to produce correct answers under reinforcement learning. This reflects the effectiveness of our reward design in guiding the model toward better reasoning outcomes.

In Figure 5(b), the temporal reward  $r_t$  (scaled to 0~1 for visibility) also demonstrates a steady increase. This suggests that the model is progressively adopting more temporally grounded reasoning strategies during training, as T-GRPO encourages preference for responses that perform better when given temporally ordered inputs. The rising curve validates the role of T-GRPO in instilling temporal awareness into the model’s reasoning process.

Interestingly, the response length curve in Figure 5(c) first drops at the beginning of RL training, then gradually increases, and finally stabilizes around a fixed length. We hypothesize that this behavior

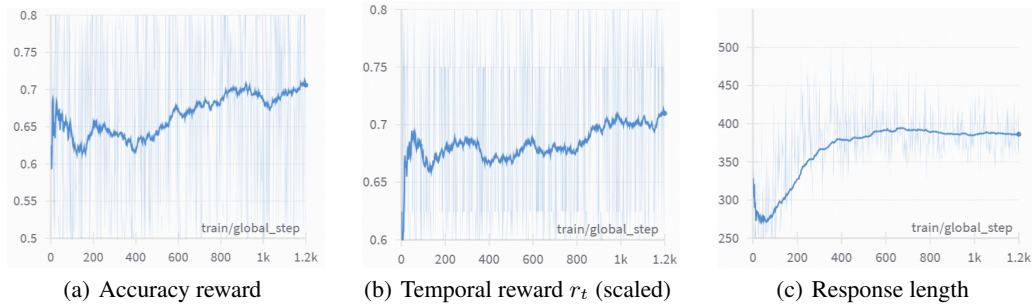


Figure 5: RL training curves.

reflects a learning transition: the model initially discards its previous, potentially sub-optimal reasoning style from supervised fine-tuning, enters a brief exploration phase with shorter outputs, and eventually converges to a new, stable reasoning policy that balances depth and conciseness. This final stabilization suggests that the model has internalized a consistent reasoning patterns that is both effective and reward-maximizing.

## 5 Conclusions, Limitations and Future Works

In this work, we present Video-R1, a reinforcement learning framework designed to enhance video reasoning capabilities in multimodal large language models (MLLMs). Motivated by the success of DeepSeek-R1 in the textual domain, we extend rule-based RL to the video setting, while addressing two key challenges: the lack of temporal inductive bias in existing RL algorithms and the scarcity of high-quality video reasoning data. To this end, we propose T-GRPO, a temporal-aware extension of GRPO that explicitly encourages the model to leverage frame order during reasoning. In parallel, we construct two curated datasets—Video-R1-COT-165k for supervised fine-tuning and Video-R1-260k for reinforcement learning—incorporating both video and image-based reasoning tasks to support stable and effective training. Experimental results across six benchmarks validate the effectiveness of our approach. We hope this work provides a foundation for further research in video reasoning with MLLMs. Some limitations and potential future works are listed as follows:

- **Increasing Frames Number.** Currently, our model is trained with 16 video frames, which may limit its ability to handle long-range temporal dependencies. In future work, we can develop more efficient training and inference strategies that allow scaling to longer videos, enabling more comprehensive temporal reasoning.
- **Better Temporal Modeling Method.** Although T-GRPO introduces effective temporal-aware reasoning, it brings additional computational overhead due to contrastive evaluation and reward calculation. This could be mitigated through inference acceleration framework such as vLLM, or by exploring more efficient mechanisms for temporal modeling.
- **Dynamic Response Length Control.** Our current length control mechanism applies a fixed reward within a predefined range, regardless of the complexity of each sample. Future work could explore dynamic length control strategies, where the model adaptively determines the appropriate response length based on the difficulty or type of the question.
- **Large-scale Reinforcement Learning.** Due to computational constraints, our current RL phase is trained with only 1k steps. Despite the promising results, we believe that increasing the RL training scale would allow the model to better explore optimal reasoning trajectories and further enhance its generalization.
- **Refined Image-to-Video Knowledge Transfer.** At present, we incorporate image-based reasoning data in a straightforward manner by mixing it into the training set. Future research could design more principled approaches for leveraging image data to more effectively transfer reasoning ability from images to videos.

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.
- [3] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.
- [4] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- [5] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*, 2024.
- [6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [7] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model, 2025.
- [8] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.
- [9] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [10] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [11] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [13] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.
- [14] Wendi Li and Yixuan Li. Process reward model with q-value rankings. *arXiv preprint arXiv:2410.11287*, 2024.
- [15] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024.
- [16] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024.

- [17] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024.
- [18] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.
- [19] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [20] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*, 2023.
- [21] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.
- [22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [23] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115. Springer, 2024.
- [24] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.
- [25] Chris Yi Peng et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought, 2025.
- [26] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025.
- [27] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024.
- [28] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.
- [29] Xiaoying Zhang, Da Peng, Yipeng Zhang, Zonghao Guo, Chengyue Wu, Chi Chen, Wei Ke, Helen Meng, and Maosong Sun. Towards self-improving systematic cognition for next-generation foundation mllms. *arXiv preprint arXiv:2503.12303*, 2025.
- [30] Yipeng Zhang, Yifan Liu, Zonghao Guo, Yidan Zhang, Xuesong Yang, Chi Chen, Jun Song, Bo Zheng, Yuan Yao, Zhiyuan Liu, et al. Llava-uhd v2: an mllm integrating high-resolution feature pyramid via hierarchical window transformer. *arXiv preprint arXiv:2412.13871*, 2024.
- [31] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.
- [32] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025.