VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning

Xinhao Li[2,1*], **Ziang Yan**[3,1*], **Desen Meng**[2], **Lu Dong**[4,1], **Xiangyu Zeng**[2,1], **Yinan He**[1]
Yali Wang[6,1], **Yu Qiao**[1,6], **Yi Wang**[†1,5], **Limin Wang**[†2,1]
[1]Shanghai AI Laboratory [2]Nanjing University [3]Zhejiang University
[4]University of Science and Technology of China [5]Shanghai Innovation Institute
[6]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
https://github.com/OpenGVLab/VideoChat-R1

Abstract

Recent advancements in reinforcement learning have significantly advanced the reasoning capabilities of multimodal large language models (MLLMs). While approaches such as Group Relative Policy Optimization (GRPO) and rule-based reward mechanisms demonstrate promise in text and image domains, their application to video understanding remains limited. This paper presents a systematic exploration of Reinforcement Fine-Tuning (RFT) with GRPO for video MLLMs, aiming to enhance spatio-temporal perception while maintaining general capabilities. Our experiments reveal that RFT is highly data-efficient for task-specific improvements. Through multi-task RFT on spatio-temporal perception objectives with limited samples, we develop **VideoChat-R1**, a powerful video MLLM that achieves state-of-the-art performance on spatio-temporal perception tasks without sacrificing chat ability, while exhibiting emerging spatio-temporal reasoning abilities. Compared to Qwen2.5-VL-7B, VideoChat-R1 boosts performance several-fold in tasks like temporal grounding (**+31.8**) and object tracking (**+31.2**). Additionally, it significantly improves on general QA benchmarks such as VideoMME (**+0.9**), MVBench (**+1.0**), and Perception Test (**+0.9**). Our findings underscore the potential of RFT for specialized task enhancement of Video MLLMs. We hope our work offers valuable insights for future RL research in video MLLMs.

1 Introduction

Recent advancements in the application of reinforcement learning (RL) in the domain of large language models (LLMs) have demonstrated remarkable progress. As evidenced by OpenAI-O1 [13], the implementation of test-time scaling strategies has shown substantial potential in enhancing LLMs' capacity for complex reasoning. Subsequently, DeepSeek-R1-Zero [10] revealed that even without extensive supervised fine-tuning, the strategic application of a rule-based reward system for reward modeling could effectively harness reinforcement learning to unlock exceptional reasoning and cognitive capabilities in language models.

Current research endeavors have increasingly focused on replicating DeepSeek-R1's success in multimodal large language models (MLLMs). Notably, Virgo [5] attempted to impart visual reasoning capabilities through knowledge distillation from open-source reasoning models such as DeepSeek-R1 [10], QwQ [28], and QvQ [27]. However, the predominant research direction [46, 19, 41, 3, 23, 20, 38, 43, 42, 4] emphasizes direct implementation of DeepSeek-R1's core Group Relative Policy Optimization (GRPO) combined with its rule-based reward system to enable visual reasoning in

* Equal contribution. † Corresponding authors.

Preprint. Under review.

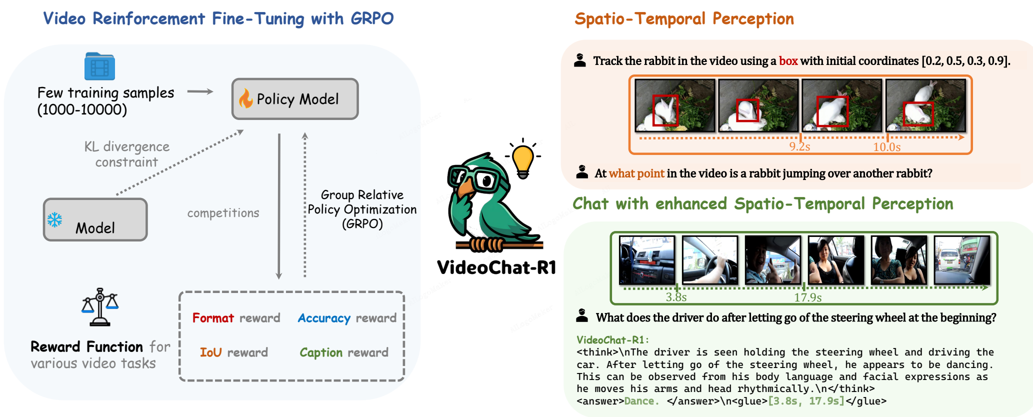


Figure 1: **Overview of VideoChat-R1.** Through reinforcement learning fine-tuning using GRPO, VideoChat-R1 has powerful spatio-temporal perception capabilities and can apply these capabilities in chatting scenarios.

MLLMs. This approach has primarily concentrated on enhancing performance in multimodal tasks involving mathematical reasoning with visual inputs and spatial localization challenges.

For video understanding, from the perspective of stimulating and evaluating reasoning abilities, there is no training and evaluation corpus as suitable as math problems and coding problems in the fields of text and images. Some works [45, 31, 6] conducted concurrently with ours have validated the superiority of the GRPO algorithm over supervised fine-tuning in some specific video tasks, such as temporal grounding and video question answer. However, deeper analysis and comprehensive ablation experiments focusing on video reasoning mechanisms remain underexplored. Current research gaps include systematic evaluations of the algorithm's generalizability across diverse video-based reasoning scenarios and granular investigations into the interplay between rule-based reward systems and multimodal temporal dependencies.

Compared with video reasoning, spatio-temporal perception is a direction in which it is easier to obtain training corpora and design a rule-based reward system. Taking the enhancement of the spatiotemporal perception ability of existing video MLLMs as the core, this paper systematically and comprehensively examines the effects of Reinforcement Fine-Tuning (RFT) on various video tasks, aiming to provide critical insights for future research. Our main findings are as follows.

- ***Reinforcement fine-tuning is data-efficient for enhancing models on specific tasks without sacrificing original capabilities***. With a small amount of data, training via RFT can yield a remarkable improvement in spatio-temporal perception ability, and there is negligible impact on the performance of out-domain tasks and the original general capabilities of the model, which outperforms traditional supervised fine-tuning significantly.
- ***Through joint reinforcement fine-tuning on multiple spatio-temporal perception tasks***, we construct **VideoChat-R1**, a powerful Video MLLM that boasts state-of-the-art spatiotemporal perception capabilities while also taking into account chat abilities. We have also discovered that training on spatio-temporal perception tasks has slightly strengthened the model's spatio-temporal reasoning abilities. Compared with Qwen2.5-VL-7B, VideoChat-R1 achieves several times the performance improvement in spatiotemporal perception tasks such as temporal grounding (**+31.8**) and object track (**+31.2**). At the same time, it also achieves significant improvements on general QA benchmarks, such as VideoMME (**+0.9**), MVBench (**+1.0**), and Perception Test (**+0.9**)

2 Related work

Reinforcement Learning Enhancement for MLLMs. Recently, works like OpenAI-o1 [13] and DeepSeek-R1 [10] have made significant breakthroughs in lifting the reasoning capabilities of large language models (LLMs) through reinforcement learning (RL). These advancements [25, 10, 26] enhance their proficiency in solving complex tasks in chains, including challenging math and coding problems. For MLLMs, many efforts [46, 19, 41, 3, 23, 20, 38, 43, 42, 4] have applied RL techniques with verifiable reward mechanisms to boost visual reasoning performance. Researches in video domain remain relatively underexplored, with only a few studies [31, 45, 6] investigating how to adapt RL-based strategies to spatiotemporal reasoning. Specifically, TimeZero [31] and R1-Omini [45] respectively demonstrate the potential of GRPO in temporal grounding and sentiment analysis. Video-R1 [6] extends GRPO to facilitate implicit temporal reasoning and achieves improvements in video spatial reasoning.

Spatio-Temporal Perception of MLLMs. Spatio-temporal perception is one of the most core capabilities of video understanding models. Despite the significant progress that Video Multimodal Large Language Models (video MLLMs) [15, 21, 16, 32, 44, 33, 18, 12, 1] have recently made in general understanding tasks such as video question answering and captioning, MLLMs' video performance still lag behind humans (event classical vision expert models) notably. Merlin [39] and TimeSuite [40] introduce spatio-temporal data augmentation for MLLM's temporal abilities, at the cost of general performance. VideoChat-TPO [36] enhances fine-grained spatio-temporal perception in videos by introducing task-specific heads using substantial training costs.

3 Methodology

We first briefly review the Group Relative Policy Optimization (GRPO) [25]. Then, we demonstrate how we design and leverage the spatio-temporal rewards for GRPO to enhance video MLLMs.

3.1 Preliminary of Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) [25] is a variant of Proximal Policy Optimization (PPO) [24] in reinforcement learning. By comparing groups of candidates responses directly, GRPO eliminates dependency on a critic model and significantly lowers training resources. Given an input question q, GRPO first generates G distinct candidate responses $o = \{o_1, \ldots, o_G\}$ through policy sampling. The MLLM serves as the reward function to get the corresponding scores $\{r_1, \ldots, r_G\}$. GRPO computes their mean and standard deviation for normalization and determines the quality of these responses:

$$A_i = \frac{r_i - \mathrm{mean}(\{r_i\}_{i=1}^{G})}{\mathrm{std}(\{r_i\}_{i=1}^{G})}, \tag{1}$$

where A_i represents the relative quality of the i-th answer. GRPO encourages the model to favor better answers with a high score within the group. The final training objective also considers preventing the optimized policy π_θ from deviating far from the original MLLM parameters π_{ref} by adding a KL-divergence term$\mathrm{D}_{\mathrm{KL}}(\cdot|\cdot)$ to :

$$\max_{\pi_\theta} \mathbb{E}_{o\sim\pi_{\theta_{\mathrm{old}}}(p)} \Big[\sum_{i=1}^{G} \frac{\pi_\theta(o_i)}{\pi_{\theta_{\mathrm{old}}}(o_i)} \cdot A_i - \beta\, \mathrm{D}_{\mathrm{KL}}\Big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\Big)\Big], \tag{2}$$

where β is a regularization coefficient, preventing excessive deviation from the reference policy during optimization.

3.2 Spatio-Temporal Rewards of Video MLLM in GRPO

We explore how to use GRPO to enhance the performance of Video MLLM in video-language understanding. We consider the five most common types of video related tasks: temporal grounding, object tracking, video question answering, captioning, and quality assessment in our experiments.

Format reward. To enable the model to output responses in the format we desire. For example, we expect the model to enclose its thought process with <think>...</think> and the answer with <answer>...</answer>, we designed a format reward R_{format} for each task. We use regular expression matching to determine whether the model adheres to our specified format:

$$R_{\mathrm{format}} = \begin{cases} 0, & \text{if output matches format,} \\ 1, & \text{if output doesn't match format.} \end{cases} \tag{3}$$

IoU reward in spatio-temporal perception. For the spatio-temporal perception such as temporal grounding and object tracking, it requires the Video MLLM to output the time interval in the video that is associated with the content of a given textual query. Evidently, we can use the Intersection over Union (IoU) between the predicted interval by the model and the ground-truth interval as the reward function. This reward function effectively characterizes the accuracy of the interval predicted by the model.

$$R_{\text{IoU}} = \frac{|\mathcal{I}_{\text{pred}} \cap \mathcal{I}_{\text{gt}}|}{|\mathcal{I}_{\text{pred}} \cup \mathcal{I}_{\text{gt}}|}, \tag{4}$$

where $\mathcal{I}_{\text{pred}}$ and $\mathcal{I}_{\text{gt}}$ are the predicted and the ground truth of time intervals or detection boxes, respectively.

Accuracy reward in classification. Discriminative tasks, such as multiple-choice video question answering and classification, aim to determine whether the model's prediction is consistent with the answer to the question. Therefore, we define:

$$R_{\text{accuacy}} = \begin{cases} 0, & \text{if } A_{\text{pred}} \neq A_{\text{gt}} \\ 1, & \text{if } A_{\text{pred}} = A_{\text{gt}}, \end{cases} \tag{5}$$

where A_{pred} and A_{gt} denote the predicted and the ground truth answers, respectively.

Recall reward in video captioning. For tasks like video captioning with open-ended outputs, it is impossible to simply compare and determine the gap between the generated caption and the ground truth caption. Therefore, we use a LLM as a "judge" to provide a reward score. In order to reduce the uncertainty of the evaluation criteria for the LLM, we first make the LLM decompose the ground truth and predicted captions into events list. Specifically, we utilize Qwen2.5-72B [37] to extract the events in the description and judge whether the events in a ground truth description can be entailed by the description predicted by the model. We calculate the event recall score as the ratio of events in a ground truth description that are entailed by the predicted description, and set different rewards according to the event recall score:

$$R_{\text{recall}} = \text{Recall}_{\text{event}}(C_{\text{pred}}, C_{\text{gt}}), \tag{6}$$

where C_{pred} and C_{gt} represent the predicted and the ground truth captions, respectively.

By combining the above reward functions, we explored how to utilize GRPO to enhance the performance of Video MLLM in various tasks. The specific details can be found in the Section 4.

3.3 Enhance Spatio-Temporal Perception of Video MLLM through GRPO

Reward fuction. We adopt different combinations of reward functions for training in different tasks. Specifically, for the temporal grounding and object tracking task, $R_{\text{st}} = R_{\text{format}} + R_{\text{IoU}}$. For the multi-choice QA and video quality assessment, $R_{\text{qa}} = R_{\text{format}} + R_{\text{accuacy}}$. For the multi-choice QA with glue (e.g. Grounding QA), $R_{\text{gqa}} = R_{\text{format}} + R_{\text{IoU}} + R_{Acc}$. For the video caption, $R_{\text{cap}} = R_{\text{format}} + R_{\text{Caption}}$.

Training data. For the temporal grounding task, we use the training set of Charade - STA [8] (5,338 samples) for training. For the object tracking task, training is conducted on the GoT - 10k [11] dataset, which has 9,335 samples. For the QA and grounding QA tasks, the validation set of NExTGQA [34] (3,358 samples) is used for training. For video captioning, FIBER-1k [35] (1,000 samples) is adopted for training. For video quality assessment, we use the quality assessment task from VidTAB [17] under the 100-shot setting, with 200 samples for training. Finally, for the training of VideoChat-R1, we perform joint training on three spatio-temporal perception-related tasks: temporal grounding, object tracking, and grounding QA. In total, 18,031 samples are used for training.

4 Experiments

Implementation details. The main experiments are all conducted based on Qwen2.5-VL-7B [1] (except for the video captioning, for which Qwen2-VL-7B [30] is used).

Benchmarks. We employ MVBench [16], Perception Test [22], VideoMME [7] for evaluation of general video understanding. Given that the majority of videos in our training set are short-length, we only use the short subset of VideoMME in testing. For the temporal grounding task, we use the test set of Charade-STA [8] for in-domain testing and the test set of ActivityNet-Grounding [14] as out-domain test data. For the object tracking task, testing is done using the GoT-10k [11] dataset. For the QA and grounding QA tasks, the test set of NExTGQA [34] is used for testing. And we use Dream-1k [29] and VidTAB-QA [17] for the video captioning and video quality access.

4.1 Evaluation of VideoChat-R1

Method	Charades-STA			ActivityNet			NExTGQA		GoT		VideoMME	MVBench	Peception Test
	mIoU	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU	acc	Overlap	R@0.5	Short-Avg	Avg	Val
Baseline													
Qwen2.5-VL-7B	29.0	24.2	11.1	21.1	15.8	7.5	15.4	59.5	12.6	1.1	71.3	66.9	69.1
SFT on specific tasks													
+SFT w/ Charades-STA	46.3	45.0	25.3	20.6	16.7	7.9	-	-	-	-	N/A*	N/A*	N/A*
+SFT w/ GoT	-	-	-	-	-	-	-	-	41.8	29.5	59.2	58.6	58.5
+SFT w/ NextQQA	-	-	-	-	-	-	28.2	64.8	-	-	60.1	59.2	60.7
GRPO on various tasks													
VideoChat-R1	**60.8**	**71.7**	**50.2**	**36.6**	**33.4**	**17.7**	32.4	**70.6**	**43.8**	**38.2**	72.2	**67.9**	**70.0**
VideoChat-R1-thinking	59.9	70.6	47.2	35.5	33.3	16.7	**36.1**	69.2	43.3	33.9	**74.2**	66.2	69.6

Table 1: **Results of VideoChat-R1 on various Video Benchmarks.** * indicates that the model has suffered from overfitting and is unable to answer the question properly. Since the number of input pixels is fixed during our evaluation, the baseline results are slightly lower than those reported in their origin paper [1].

As shown in Table 1, after training with GRPO on spatio-temporal perception datasets, both VideoChat-R1 and VideoChat-R1-thinking significantly outperform the performance of Qwen2.5-VL and that of models fine-tuned through SFT for a single specific task across various spatiotemporal perception benchmarks and the general understanding benchmark VideoMME. This validates the effectiveness of our approach, which leverages multiple spatiotemporal perception datasets and RFT for enhancing spatiotemporal perception.

Meanwhile, we observe that for spatio-temporal perception tasks, engaging in thinking processes does not necessarily lead to performance gains. However, for tasks such as QA and VideoMME, which may require complex reasoning, conducting inferences during testing can result in notable performance improvements.

4.2 Ablation Studies and Discussions

Method	Epochs	Training Prompt		Test Prompt		Charades-STA (in domain)				ActivityNet (out domain)				VideoMME
		Think	Answer	Think	Answer	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	Short-Avg
Vision Experts														
FlashVTG [2]	-	-	-	-	-	-	-	70.3	49.9	-	-	-	-	-
InternVideo2-6B [32]	-	-	-	-	-	-	-	70.0	49.0	-	-	-	-	-
SG-DETR [9]						-	-	71.1	52.8	-	-	-	-	-
MLLMs														
Qwen2.5-VL-7B	-	-	-		✓	29.0	44.7	24.2	11.1	21.1	28.3	15.8	7.4	71.3
(baseline)	-	-	-	✓	✓	28.1	41.8	23.4	11.1	17.7	22.7	13.4	7.7	71.3
+ SFT	1		✓		✓	46.3	63.9	45.0	25.3	20.6	30.2	16.7	7.9	N/A*(-71.3)
	3		✓		✓	34.6(+6.5)	51.7	36.3	20.6	17.3(-3.8)	26.1	10.0	3.9	N/A*(-71.3)
+ GRPO	1		✓		✓	58.7	80.9	67.7	45.4	31.9	46.3	28.8	14.1	72.6
	1	✓	✓	✓	✓	59.3(+31.2)	81.7	70.4	46.0	30.7(+13.0)	45.0	27.5	12.9	73.6(+2.3)
	3	✓	✓	✓	✓	61.3(+33.2)	83.1	72.8	51.5	34.3(+16.6)	50.4	32.2	16.2	70.9(-0.4)

Table 2: **Results of Temporal Grounding Task.** * indicates that the model has suffered from overfitting and is unable to answer the question properly.

Method	GoT			VideoMME
	Average overlap	R@0.5	R@0.7	Short-Avg
Qwen2.5-VL-7B	12.6	1.1	0	71.3
+SFT	41.8	29.5	3.9	59.2
+GRPO	42.5(+29.2)	30.6(+29.5)	3.9(+3.9)	71.4(+0.1)

Table 3: **Results of Object Tracking.** We use 8 frames as input for training and evaluation.

Temporal Grounding and Object tracking. As shown in Table 2 and Table 3, fine-tuning Qwen2.5-VL using GRPO significantly improves the performance of temporal grounding and object tracking tasks. Additionally, it slightly enhances the performance on the general understanding benchmark VideoMME. Even when training for more epochs, GRPO is less prone to overfitting compared to SFT. Instead, it can continuously improve the performance of temporal grounding, eventually surpassing the performance of previous expert models. Moreover, stimulating the model's thinking ability provides some benefits for both temporal grounding and VideoMME tasks.

Video Question Answer. As shown in Table 4, for the video question answering task, we selected the multi-choice QA task, which is easy to evaluate, for our experiments. Additionally, we explored the grounding QA task. In this task, when answering questions, the model is required to simultaneously provide the temporal cues on which its answers are based. Using merely a little over three thousand training data samples, we found that GRPO demonstrated remarkable fine-tuning capabilities. Not only did it lead to a substantial improvement in the performance of the NExTGQA task, but it also brought about a noticeable enhancement in the VideoMME task. We noticed that, unlike the previous strongly spatiotemporal perception tasks such as temporal grounding, thinking played a significant role in the QA task. Meanwhile, the glue signals also provided some assistance for relatively complex video understanding tasks like VideoMME.

Method	Training Prompt			Test Prompt			NExTGQA		VideoMME
	Think	Answer	Glue	Think	Answer	Glue	mIoU	acc	Short-Avg
Direct Output									
Qwen2.5-VL-7B					✓		-	41.7	71.3
(baseline)					✓	✓	15.4	59.5	-
+ SFT		✓			✓		-	65.1	60.2
		✓	✓		✓	✓	28.2(+12.8)	64.8(+5.3)	60.1(-11.2)
+ GRPO		✓			✓		-	70.1	71.7
		✓			✓	✓	16.2	70.2	71.7
		✓	✓		✓	✓	35.1(+19.7)	68.7(+9.2)	72.0(+0.7)
Chain-of-thought Output									
Qwen2.5-VL-7B				✓	✓		-	47.7	73.0
				✓	✓	✓	20.2	53.3	72.2
+ GRPO	✓	✓		✓	✓		-	68.8	74.7
	✓	✓	✓	✓	✓	✓	32.9(+12.7)	66.9(+13.6)	75.3(+3.1)

Table 4: **Results of Multi-Choice Video QA.**

Video Caption and Video Quality Assessment. For the Video Caption and Video Quality Assessment tasks, we found that GRPO still demonstrated its advantages over SFT. Surprisingly, although the reward we designed for the Caption task was not entirely rigorous and reasonable, it still enabled the model to achieve significant improvements in metrics on the Dream-1k benchmark.

Method	Dream-1k			VidTAB-QA
	F1	Precision	Recall	Accuracy
Baseline	30.6	33.8	27.9	70.7
+ SFT	31.4	32.6	30.2	71.7
+ GRPO	38.2(+7.6)	45.4(+11.6)	33.1(+5.2)	72.6(+1.9)

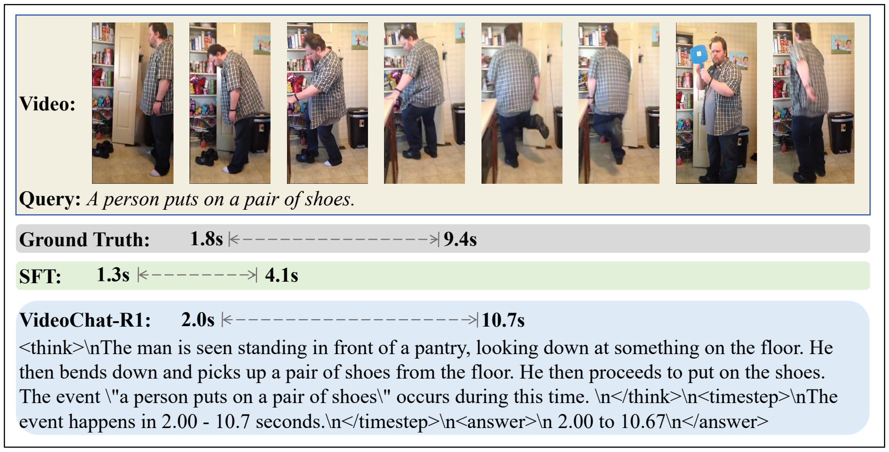
Table 5: **Results of Video Caption and Video Quality Access.**

Muti-task Co-training. As shown in Table 6, we found that mixed training of different spatiotemporal perception tasks using GRPO can yield a synergistic improvement effect. Training with the multiple tasks achieves nearly the best results across all benchmarks. This reveals the potential of GRPO for larger-scale and multi-task collaborative training in the future.

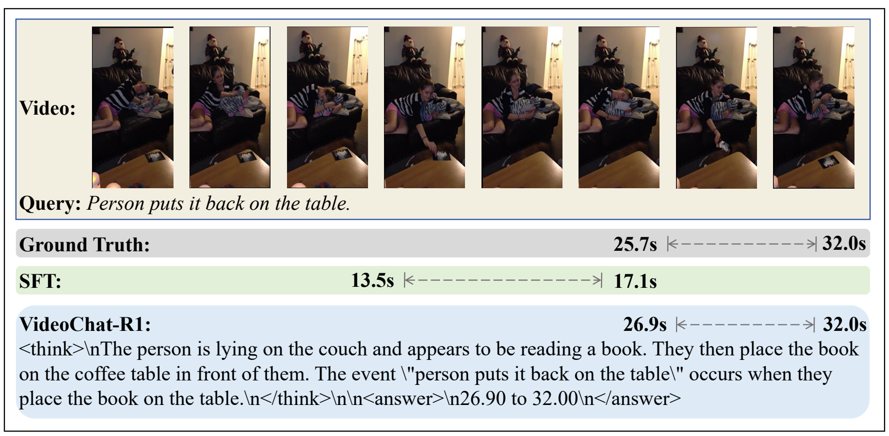
GRPO vs. SFT. It can be observed that across various types of tasks, GRPO outperforms SFT. Whether it is in terms of the performance on in-domain tasks, out-domain tasks, or the preservation

Method	Charades-STA			ANet			NExTGQA		GoT		VideoMME
	mIoU	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU	acc	Overlap	R@0.5	Short-Avg
Qwen2.5-VL-7B	29.0	24.2	11.1	21.1	15.8	7.5	15.4	59.5	12.6	1.1	71.3
+GRPO w/ STA	59.3	70.4	46.0	30.7	27.5	12.9	31.4	61.2	27.8	12.9	**72.6**
+GRPO w/GQA	36.0	33.5	15.5	24.9	20.6	10.7	35.1	68.7	36.1	26.7	72.0
+GRPO w/ GoT	28.7	25.1	9.6	20.1	16.2	6.8	15.6	60.5	42.5	30.6	71.4
+GRPO w/ STA-GQA	59.8	69.7	47.0	33.7	31.0	16.0	**35.7**	67.7	36.5	28.9	72.2
+GRPO w/ STA-GQA-GoT	**60.8**	**71.7**	**50.2**	**36.6**	**33.4**	**17.7**	32.4	**70.6**	**43.8**	**38.2**	72.2

Table 6: **Results of Cotraining on Spatio-Temporal Tasks.**



(1)



(2)

Figure 2: **Examples on temporal grounding task.** VideoChat-R1 gives a more accurate time interval after thinking.

of the original general performance, our experimental results demonstrate that GRPO is a promising fine-tuning approach. We will leave the large-scale comparison for future research.

Chain-of-thought vs. Direct Output. Based on the video tasks and experiments we have explored, which focus on spatiotemporal perception, the output of the chain of thought has not demonstrated obvious advantages. In some cases, it is even inferior to the direct output. We believe that how to define appropriate video reasoning tasks and evaluation methods remains to be explored. The existing training data is insufficient to activate the model to output truly effective video reasoning chains.

4.3 Qualitative Results

As shown in Figure 2 and 3, we provide visualizations of VideoChat-R1's outputs for the temporal grounding and video QA tasks. Note that VideoChat-R1 possesses strong spatiotemporal perception.

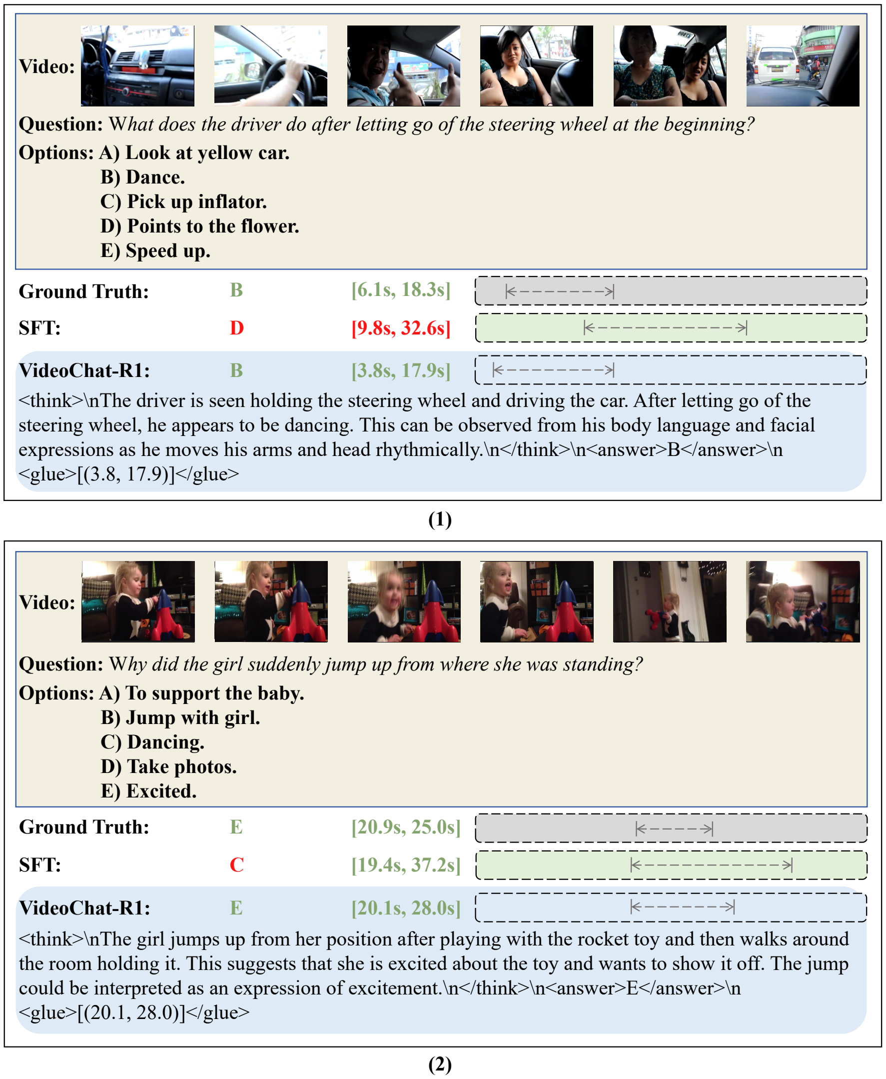


Figure 3: **Examples on Video QA task.** It can be seen that VideoChat-R1 can not only answer questions correctly but also provide relatively accurate reference time periods (glue).

Additionally, its thought process exhibits a certain degree of logic. However, due to the limitations of the tasks, compared with the thought processes for math problems in the text and image domains, the thought process of VideoChat-R1 is still rather simplistic and rudimentary.

5 Conclusions

In this work, we systematically investigate the role of reinforcement fine-tuning (RFT) with Group Relative Policy Optimization (GRPO) in enhancing video-centric multimodal large language models (MLLMs). Our experiments demonstrate that RFT is a highly data-efficient paradigm for task-specific improvements, enabling VideoChat-R1—a model trained with limited samples via multi-task RFT—to achieve state-of-the-art performance on spatio-temporal perception tasks while preserving general chat capabilities and exhibiting emergent spatiotemporal reasoning. We believe our work can present relevant insights for future research efforts in reinforcement learning of video MLLMs.

References

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

[2] Zhuo Cao, Bingqing Zhang, Heming Du, Xin Yu, Xue Li, and Sen Wang. Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. *arXiv preprint arXiv:2412.13441*, 2024.

[3] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.

[4] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.

[5] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025.

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

[7] Chaoyou Fu, Yuhan Dai, Yondong 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.

[8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[9] Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. *arXiv preprint arXiv:2410.01615*, 2024.

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

[11] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1562–1577, 2019.

[12] Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: A comprehensive benchmark and memory-augmented method. *arXiv preprint arXiv:2501.00584*, 2024.

[13] 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.

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

[15] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[16] 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.

[17] Xinhao Li, Zhenpeng Huang, Jing Wang, Kunchang Li, and Limin Wang. Videoeval: Comprehensive benchmark suite for low-cost evaluation of video foundation model. *arXiv preprint arXiv:2407.06491*, 2024.

[18] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

[19] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[20] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[21] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[22] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023.

[23] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[25] 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.

[26] 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.

[27] Qwen Team. Qvq: To see the world with wisdom, December 2024. URL `https://qwenlm.github.io/blog/qvq-72b-preview/`.

[28] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL `https://qwenlm.github.io/blog/qwq-32b/`.

[29] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024.

[30] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

[32] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[33] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

[34] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.

[35] Yifan Xu, Xinhao Li, Yichun Yang, Rui Huang, and Limin Wang. Fine-grained video-text retrieval: A new benchmark and method. *arXiv preprint arXiv:2501.00513*, 2024.

[36] Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, et al. Task preference optimization: Improving multimodal large language models with vision task alignment. *arXiv preprint arXiv:2412.19326*, 2024.

[37] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

[39] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. In *European Conference on Computer Vision*, pages 425–443. Springer, 2024.

[40] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

[41] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025.

[42] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025.

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

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

[45] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv e-prints*, pages arXiv–2503, 2025.

[46] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.