TinyLLaVA-Video-R1: Towards Smaller LMMs for Video Reasoning

Xingjian Zhang[1,*] **Siwei Wen**[1,2,*] **Wenjun Wu**[1,2,3] **Lei Huang**[1,2,3,✉]

[1]SKLCCSE, Institute of Artificial Intelligence, Beihang University, Beijing, China
[2]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beihang University
[3]Hangzhou International Innovation Institute, Beihang University, Hangzhou, China

{huangleiai}@buaa.edu.cn

Abstract

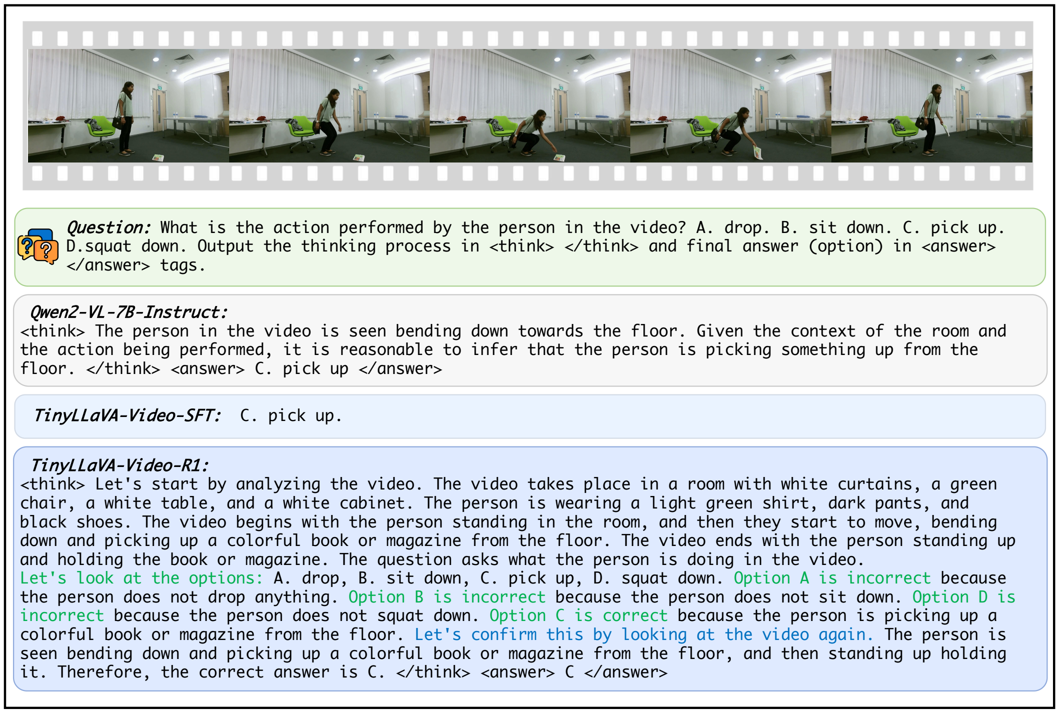
Recently, improving the reasoning ability of large multimodal models (LMMs) through reinforcement learning has made great progress. However, most existing works are based on highly reasoning-intensive datasets such as mathematics and code, and researchers generally choose large-scale models as the foundation. We argue that exploring small-scale models' reasoning capabilities remains valuable for researchers with limited computational resources. Moreover, enabling models to explain their reasoning processes on general question-answering datasets is equally meaningful. Therefore, we present the small-scale video reasoning model TinyLLaVA-Video-R1. Based on TinyLLaVA-Video [27], a traceably trained video understanding model with no more than 4B parameters, it not only demonstrates significantly improved reasoning and thinking capabilities after using reinforcement learning on general Video-QA datasets, but also exhibits the emergent characteristic of ”aha moments”. Furthermore, we share a series of experimental findings, aiming to provide practical insights for future exploration of video reasoning (thinking) abilities in small-scale models. It is available at https://github.com/ZhangXJ199/TinyLLaVA-Video-R1.

1 Introduction

Since DeepSeek-R1 [6] demonstrated that pure reinforcement learning can significantly enhance a model's reasoning capabilities, many subsequent works [1, 7, 15, 30, 16] have also explored improving the reasoning abilities of multimodal models, achieving notable progress. Most of these efforts focus on extending reasoning capabilities to the image modality [15, 1], conducting research using strong reasoning data such as math-image pairs and spatial reasoning [16, 7, 30], or task-specific data like grounding [14, 18]. However, existing research on video reasoning models has not made significant progress due to the scarcity of highly reasoning-intensive data.

Open-R1-Video [21] is the first to introduce reasoning into the video domain, however, its performance on general Video-QA datasets is unsatisfactory, with benchmark results even showing a decline. Subsequently, Video-R1 [3] successfully integrates strong reasoning image-text pairs for video reasoning, achieving remarkable performance, but they make preliminary attempts and argue that small-scale models cannot produce effective reasoning processes. However, the high computational cost of large-scale models remains a significant barrier for many researchers with limited resources. Therefore, exploring the reasoning capabilities of smaller models is still necessary.

[*]denotes equal contributor; ✉ denotes corresponding author.



Question: What is the action performed by the person in the video? A. drop. B. sit down. C. pick up. D.squat down. Output the thinking process in <think> </think> and final answer (option) in <answer> </answer> tags.

Qwen2-VL-7B-Instruct:
<think> The person in the video is seen bending down towards the floor. Given the context of the room and the action being performed, it is reasonable to infer that the person is picking something up from the floor. </think> <answer> C. pick up </answer>

TinyLLaVA-Video-SFT: C. pick up.

TinyLLaVA-Video-R1:
<think> Let's start by analyzing the video. The video takes place in a room with white curtains, a green chair, a white table, and a white cabinet. The person is wearing a light green shirt, dark pants, and black shoes. The video begins with the person standing in the room, and then they start to move, bending down and picking up a colorful book or magazine from the floor. The video ends with the person standing up and holding the book or magazine. The question asks what the person is doing in the video. Let's look at the options: A. drop, B. sit down, C. pick up, D. squat down. Option A is incorrect because the person does not drop anything. Option B is incorrect because the person does not sit down. Option D is incorrect because the person does not squat down. Option C is correct because the person is picking up a colorful book or magazine from the floor. Let's confirm this by looking at the video again. The person is seen bending down and picking up a colorful book or magazine from the floor, and then standing up holding it. Therefore, the correct answer is C. </think> <answer> C </answer>

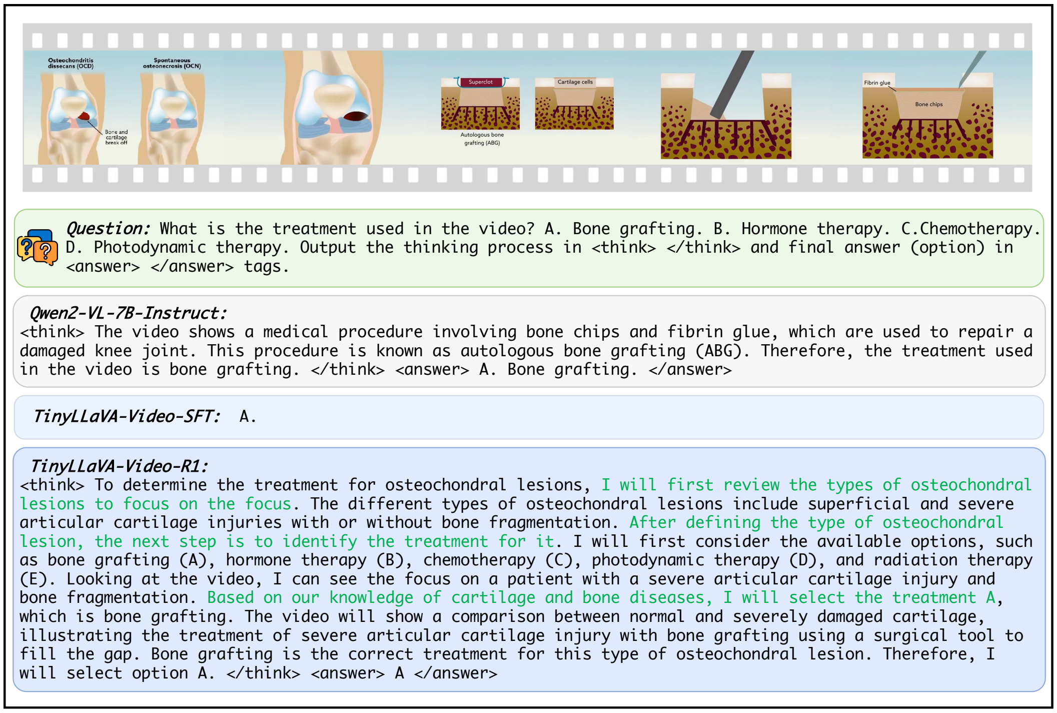
Figure 1: A case of TinyLLaVA-Video-R1 on video understanding data, sourced from MVBench. The model demonstrates the ability to perceive video scenes and analyze options, while also exhibiting reflective and backtracking behavior (highlighted in blue).

In this work, we propose the small-scale video reasoning model TinyLLaVA-Video-R1, based on the traceably trained model TinyLLaVA-Video [27]. After reinforcement learning on general Video-QA datasets, the model not only significantly improves its reasoning and thinking abilities, but also exhibits the emergent characteristic of “aha moments”, which is more meaningful than simply generating answers through perception. Moreover, through extensive experiments under various configurations, we have obtained a series of insightful findings. We believe these discoveries will provide valuable guidance for future exploration of video reasoning capabilities in small-scale models.

2 Related Work

Large Language Reasoning Models. Recently, Kimi K1.5 [19] and DeepSeek-R1 [6] have gained widespread attention for their significant improvements in model performance on reasoning tasks through using reinforcement learning. Unlike approaches relying on process-supervised reward models [5], the Group Relative Policy Optimization (GRPO) algorithm [17] relying on rule-based rewards not only substantially reduces computational costs but has also sparked a new wave of interest due to intriguing phenomena such as ”aha moments” observed during training. Under the influence of this research paradigm, numerous follow-up studies have advanced along this technical path, some have successfully replicated the method and extended it to multimodal domains [1, 7, 30], achieving notable progress, while others [13, 24] have focused on algorithmic optimization to further enhance the reasoning efficiency and performance of models.

Multimodal Reasoning Models. With the remarkable advancement in reasoning capabilities of language models, an increasing number of studies have attempted to extend them to multimodal domains. LMM-R1 [16] proposes a two-stage training strategy to enhance the reasoning performance of multimodal models; Vision-R1 [7] attempts to address the post-cold-start overthinking issue in multimodal models; Video-R1 [3] develops T-GRPO to further enhance model’s video comprehension. While these studies have made notable progress, most tend to adopt base models with 7B or more parameters to ensure superior reasoning effectiveness and robust performance.


Osteochondritis dissecans (OCD)
Spontaneous osteonecrosis (OCN)
Bone and cartilage break off
Superclot
Cartilage cells
Autologous bone grafting (ABG)
Fibrin glue
Bone chips

Question: What is the treatment used in the video? A. Bone grafting. B. Hormone therapy. C.Chemotherapy. D. Photodynamic therapy. Output the thinking process in <think> </think> and final answer (option) in <answer> </answer> tags.

Qwen2-VL-7B-Instruct:
<think> The video shows a medical procedure involving bone chips and fibrin glue, which are used to repair a damaged knee joint. This procedure is known as autologous bone grafting (ABG). Therefore, the treatment used in the video is bone grafting. </think> <answer> A. Bone grafting. </answer>

TinyLLaVA-Video-SFT: A.

TinyLLaVA-Video-R1:
<think> To determine the treatment for osteochondral lesions, I will first review the types of osteochondral lesions to focus on the focus. The different types of osteochondral lesions include superficial and severe articular cartilage injuries with or without bone fragmentation. After defining the type of osteochondral lesion, the next step is to identify the treatment for it. I will first consider the available options, such as bone grafting (A), hormone therapy (B), chemotherapy (C), photodynamic therapy (D), and radiation therapy (E). Looking at the video, I can see the focus on a patient with a severe articular cartilage injury and bone fragmentation. Based on our knowledge of cartilage and bone diseases, I will select the treatment A, which is bone grafting. The video will show a comparison between normal and severely damaged cartilage, illustrating the treatment of severe articular cartilage injury with bone grafting using a surgical tool to fill the gap. Bone grafting is the correct treatment for this type of osteochondral lesion. Therefore, I will select option A. </think> <answer> A </answer>

Figure 2: A case of TinyLLaVA-Video-R1 on video reasoning data, sourced from MMVU. The model demonstrates comprehensive video content understanding and the capability to derive correct answers through analytical reasoning.

Although some preliminary work has explored small-scale multimodal models [16, 1, 30], these investigations have primarily focused on the image modality, leaving research on video reasoning significantly underdeveloped. Due to the scarcity of high-quality video reasoning datasets and the inherent capability limitations of models, small-scale video reasoning models face multiple challenges including limited response length, difficulties in presenting coherent reasoning chains, and constrained overall performance improvement [3]. These critical challenges demand thorough investigation and substantive technical breakthroughs.

3 Methods

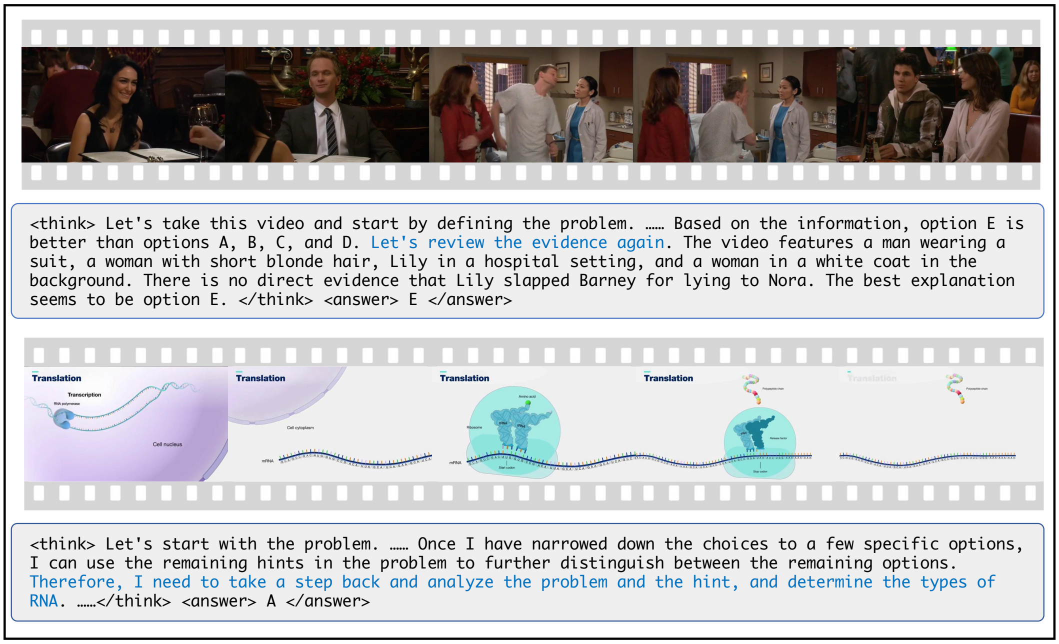
To explore the video reasoning capabilities of small-scale models, we conduct experiments on TinyLLaVA-Video [27]. We utilize the GRPO algorithm on the general Video-QA dataset NextQA and made specific modifications to the reward rules: adding a continuous length reward to the format reward and introducing penalties for incorrect answers. The experimental results in Section 4 demonstrate the effectiveness of these modifications.

3.1 TinyLLaVA-Video

TinyLLaVA-Video is a fully open-source small-scale video understanding model that employs Qwen2.5-3B [8] as its language model and SigLIP [25] as its visual encoder. It delivers competitive performance across multiple benchmarks. Crucially, its training data are fully open-sourced, and the entire training process remains traceable. This effectively prevents the repeated use of identical data across different training phases, thereby avoiding the introduction of uncontrolled variables and ensuring more reliable experimental results and conclusions. Such reproducibility and controllability represent a distinct advantage over models that only release weights, making TinyLLaVA-Video an ideal foundational model for our experiments on investigating video reasoning.

3.2 Group Relative Policy Optimization (GRPO)

We follow the GRPO algorithm [17] to train the model. For each question q, the policy model generates a set of candidate responses $\{O_1, O_2, \ldots, O_G\}$, computes the corresponding rewards



<think> Let's take this video and start by defining the problem. …… Based on the information, option E is better than options A, B, C, and D. Let's review the evidence again. The video features a man wearing a suit, a woman with short blonde hair, Lily in a hospital setting, and a woman in a white coat in the background. There is no direct evidence that Lily slapped Barney for lying to Nora. The best explanation seems to be option E. </think> <answer> E </answer>

<think> Let's start with the problem. …… Once I have narrowed down the choices to a few specific options, I can use the remaining hints in the problem to further distinguish between the remaining options. Therefore, I need to take a step back and analyze the problem and the hint, and determine the types of RNA. ……</think> <answer> A </answer>

Figure 3: Cases of ”aha moment”, where the model demonstrates reflection and backtracking during its reasoning process (highlighted in blue). The cases are from MVBench and MMVU respectively.

$\{r_1, r_2, \ldots, r_G\}$ based on the reward rules. And then these rewards are normalized to calculate the advantage for each response. Subsequently, the model is optimized through maximization of the following objective function:

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

where π_θ and $\pi_{\theta_{old}}$ are the current and old policy, ϵ and β are hyper-parameters, and A_i is the advantages defined as:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}. \tag{2}$$

In addition, our experimental observations reveal an issue analogous to DAPO [24]: when all responses in a set $\{O_i\}$ are correct and given equal rewards, their computed advantages vanish to zero. This phenomenon affects policy updates and diminishes sample efficiency. To maximize the utility of each sample, we introduce an additional gaussian noise $\mathcal{N}(0,\ 0.02^2)$ to the advantages. Although the noise induces only minor perturbations, it ensures intra-group advantage diversity across responses.

3.3 Training Data and Template.

We select multiple choice questions from the NextQA [23] subset of LLaVA-Video-178K [28] as training data. To maintain manageable training time with limited computational resources, we only choose the subset of data with a duration of 0 to 30 seconds, which contains 5,496 samples. It is a weak reasoning dataset, where the questions are more perception-oriented and exhibit weaker logical reasoning. However, we hypothesize that the model's reasoning abilities are likely predominantly derived from reinforcement learning, and we still aim to guide it to demonstrate its thought process by articulating the reasoning behind its choices, rather than merely providing an answer.

During training, for each input question, in addition to the system template, we append the following prompt at the end: `Output the thinking process in <think> </think> and final answer (option) in <answer> </answer> tags.` Moreover, when computing rewards for responses, we strictly enforce the model to adhere to this format.

Model	LLM size	Answer Type	MVBench	Video-MME (wo sub)	MLVU	MMVU (mc)
LLaMA-VID [10]	7B	Option	41.4	-	33.2	-
LLaVA-NeXT [12]	7B	Option	-	-	39.3	29.2
VideoLLaVA [11]	7B	Option	-	39.9	47.3	-
ShareGPT4Video [2]	8B	Option	-	39.9	46.4	-
LLaVA-Mini [26]	7B	Option	44.5	-	42.8	-
InternVideo2 [22]	8B	Option	-	41.9	-	39.0
TinyLLaVA-Video-SFT	3B	Option	49.0	42.2	49.2	46.1
TinyLLaVA-Video-ColdStart	3B	Reason	33.2	26.6	28.6	22.7
TinyLLaVA-Video-R1	3B	Reason	**49.5**	**46.6**	**52.4**	**46.9**

Table 1: The performance of TinyLLaVA-Video-R1 on multiple benchmarks. ”Option” indicates that the model only needs to answer with the selected choice, while ”Reason” means the model must output both the answer and the reasoning process according to the format requirements. Here, MMVU is categorized as a video reasoning benchmark, the remaining benchmarks are designed for general-purpose video evaluation. The best results are indicated by **boldface**.

3.4 Reward Rules.

We also avoid using a reward model and define reward rules based on the format and accuracy of the responses as follows:

Format reward. We require the thought process to be enclosed within `<think></think>`, and the final answer to be enclosed within `<answer></answer>`. These four tags can appear only once in the entire response, and if followed, the model will receive a format reward $FR = r_0 + LR$. Here, r_0 represents the base reward for adhering to the required response format, and LR is the continuous length reward designed to encourage the model to generate longer outputs, calculated as:

$$LR = \min\left(1, \frac{Len}{ML}\right) \times r_1. \tag{3}$$

Here, Len represents the length of the response extracted from within the `<think></think>` tags, and ML represents the maximum length corresponding to the upper limit of the reward. In our experiments, we set $r_0 = r_1 = 0.5$, thus the format reward is limited to a maximum of 1.

Accuracy reward. We design the accuracy reward AR based on the answer. We extract the final answer from `<answer></answer>` and compare it with the label. The model will receive an accuracy reward of $AR = r_2 > 0$, if the answer is correct. Responses with either format errors preventing answer extraction or incorrect answers will result in zero accuracy reward, i.e. $AR = 0$. To ensure that the accuracy reward and the format reward have equal importance, we set $r_2 = r_0 + r_1$ in our experiments.

To encourage the model to increase the response length only when answering correctly, rather than arbitrarily increasing the length at the cost of accuracy, we deviate from most existing approaches that simply define the total reward as the sum of format reward and accuracy reward. Instead, we introduce a penalty for incorrect answers, with the total reward R defined by the following formula:

$$R = \begin{cases} AR + FR, & \text{if } FR > 0 \text{ and } AR = r_2 \\ -FR, & \text{if } FR > 0 \text{ and } AR = 0 \\ -(r_0 + r_1 + r_2), & \text{if } FR = 0 \end{cases} \tag{4}$$

When the model’s answer is correct, the longer the reasoning process, the higher the reward. In contrast, if the answer is incorrect, the longer the reasoning process, the higher the penalty incurred.

4 Experiments

4.1 Experimental Settings

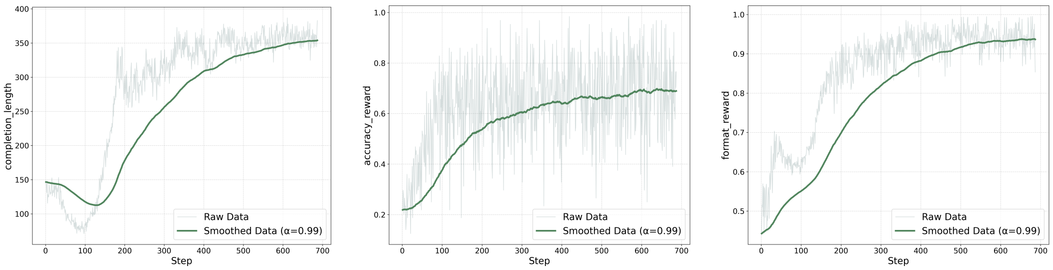
We conduct experiments on 8 NVIDIA A100-40G GPUs. During training, we keep the vision encoder frozen and update the connector and language model. We set the learning rate at 1e-6 for stable training.

To facilitate rapid adaptation to reasoning format and ensure training stability, we first finetune the model using 16 human-annotated cold-start samples, resulting in TinyLLaVA-Video-ColdStart. We then adopt it as the base model for reinforcement learning and train on 5,496 NextQA data for one epoch to obtain TinyLLaVA-Video-R1.

For evaluation, we select four commonly used video understanding and reasoning benchmarks: MVBench [9], VideoMME[4], MLVU [31], and MMVU [29]. These benchmarks encompass videos from multiple disciplines and domains, with a wide range of durations, enabling a comprehensive assessment of the model's capabilities.

4.2 Main Results and Aha Moment

As shown in Figure 4, during training, both the response length and rewards demonstrate stable growth. As presented in Table 1, compared to TinyLLaVA-Video-SFT, which is trained on the same dataset using supervised learning, TinyLLaVA-Video-R1 shows superior performance across multiple benchmarks. Additionally, compared to the base model TinyLLaVA-Video-ColdStart, TinyLLaVA-Video-R1 not only adheres to the required response format but also demonstrates improved reasoning capabilities.



(a) Evolution in completion length. (b) Evolution in accuracy reward. (c) Evolution in format reward.

Figure 4: Evolution in key metrics during the training of TinyLLaVA-Video-R1. Under our reward rule settings, both the response length and rewards of TinyLLaVA-Video-R1 gradually increased during training.

As shown in Figures 1 and 2, we further illustrate the model's reasoning ability. The model can comprehend and analyze video content, evaluate each option step-by-step, and ultimately provide an answer. Compared to models that only output answers without reasoning, TinyLLaVA-Video-R1 generates meaningful thought processes, making its responses more interpretable and valuable. This represents a significant advantage of video reasoning models over conventional video understanding models.

Similar to other works that use reinforcement learning to enhance model reasoning capabilities, we also reproduce the ”aha moment” in TinyLLaVA-Video-R1, where the model exhibits emergent behaviors such as self-verification during its reasoning process. Our experimental results confirm that even when trained with weakly-reasoned general video data through reinforcement learning, the smaller model can still demonstrate retrospection and reflection.

As highlighted in the blue annotations in Figures 1 and 3, the model revisits and verifies its initial reasoning after completing a round of thought. This behavior indicates that the model does not merely perform perception but also engages in continuous thinking and self-checking.

4.3 Ablation Study

In this section, we present ablation studies on methods and key experimental findings that contribute significantly to the performance enhancement of TinyLLaVA-Video-R1.

4.3.1 Impact of Cold-Start Data

Due to the limitations of language models, when we directly use TinyLLaVA-Video as the base model without length reward, we find that as training progresses, the model has a certain probability of learning to 'take shortcuts'. While adhering to the required format, all responses omit the reasoning process and are structured strictly as `<think> </think> <answer> option </answer>`. We observe similar experimental phenomena when conducting experiments on Qwen2-VL-2B [20], so we believe this is a common issue with small-scale models.

However, when we perform a cold start with 16 human-annotated CoT data, this phenomenon no longer appear during the experiments. At the same time, the model also learn to comply with the format requirements more quickly. Therefore, we believe that cold starting is necessary for reasoning in small-scale models. Even a small amount of cold start data can be very helpful for stabilizing model training.

4.3.2 Impact of Refinement of Format Rewards

In our initial experiments, similar to other works, we only apply format reward without incorporating continuous length reward. However, constrained by the capabilities of small-scale language models, training under this setup does not lead to an increase in response length, and even results in a slight decline. After introducing continuous length reward, the model's response length significantly increases during training, as shown in Figure 5. However, we observe that under this setup, the model engages in some meaningless reasoning to increase response length, which does not improve performance and even leads to a significant increase in training time. When incorporating answer correctness penalty into the total reward as described in Section 3.4, we observe both qualitative improvements in model responses and continued growth in output length and rewards throughout training as shown in Figure 4.

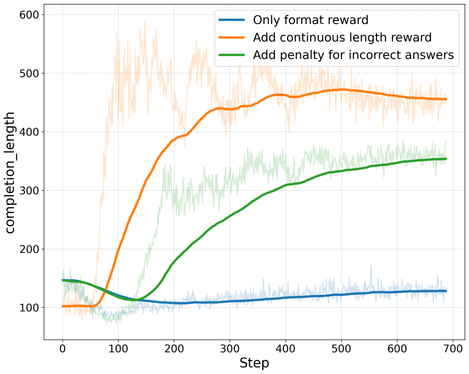


Figure 5: The variation in response length during training under different settings.

4.3.3 Other Experimental Explorations

Meanwhile, we also experiment with some existing improvements to GRPO. Some studies [24, 13] suggest that the distribution of reasoning models may differ significantly from the initial model, so removing the KL divergence can eliminate constraints on the model. As shown in Figure 6, our experiments similarly demonstrate that eliminating the KL divergence improves model performance. Additionally, Dr. GRPO [13] argues that the increase in response length may also stem from inherent biases in the GRPO objective function. After removing the KL divergence, we further exclude the response length term from the objective function and the reward variance term from the advantage calculation. As shown in Figure 6, the performance of the model improves again. At the same time, we observe a noticeable reduction in response length, the model tends to only provide descriptions of the video content while omitting analysis of the answer. We attribute this to the lack of strong reasoning in the training dataset, which fails to stimulate deep logical reasoning in the models.

5 Conclusion and Future Work

In this work, we propose the small-scale video reasoning model TinyLLaVA-Video-R1, which is trained using reinforcement learning on a general Video-QA dataset. It not only significantly enhances reasoning and thinking capabilities, but also exhibits the emergent characteristic of "aha

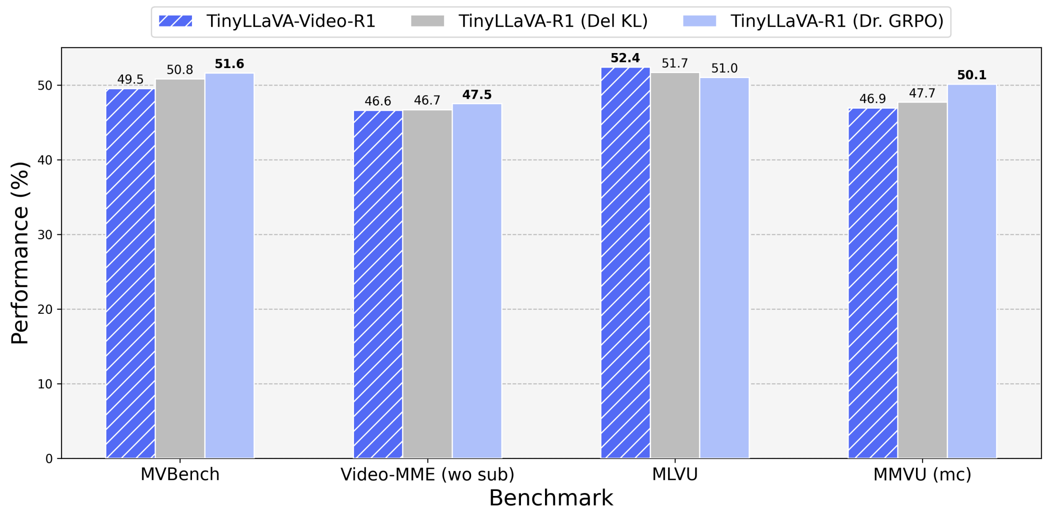


Figure 6: Ablation study on TinyLLaVA-R1 variants across multiple benchmarks. We compare the original TinyLLaVA-Video-R1 with two ablated versions: removing the KL divergence term (Del KL) and replacing the original GRPO with Dr. GRPO. Results are reported on MVBench, Video-MME (without subtitle input), MLVU, and MMVU (multiple-choice subset). **Bold** values indicate the best performance for each benchmark.

moment". Additionally, we present a series of experimental findings, hoping this work will provide valuable insights for future practitioners exploring the video reasoning abilities of small-scale models. We will further investigate small-scale video reasoning models, with potential future directions as follows:

- **Introducing high-quality video reasoning data.** Currently, TinyLLaVA-Video-R1 is trained only on general video question-answering data. We aim to explore the upper limits of the model's reasoning capabilities by introducing higher-quality video reasoning data.
- **Improving reinforcement learning algorithms.** Currently, TinyLLaVA-Video-R1 employs the GRPO algorithm for training. However, this approach exhibits notable limitations. To enhance its effectiveness in video reasoning tasks, we plan to refine the algorithm by addressing the key challenges observed in our experiment.

Acknowledgment. This work was partially supported by the National Science and Technology Major Project (Grant No. 2022ZD0116310), National Natural Science Foundation of China (Grant No. 62476016), the Fundamental Research Funds for the Central Universities.

References

[1] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02. 1, 2, 3

[2] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 5

[3] 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. 1, 2, 3

[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. 6

[5] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025. 2

[6] 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. 1, 2

[7] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 1, 2

[8] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024. 3

[9] 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. 6

[10] 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, 2025. 5

[11] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 5

[12] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5

[13] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. 2, 7

[14] 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. 1

[15] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 1

[16] 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. 1, 2, 3

[17] 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. 2, 3

[18] Haozhan Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. `https://github.com/om-ai-lab/VLM-R1`, 2025. Accessed: 2025-02-15. 1

[19] 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. 2

[20] 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. 7

[21] Xiaodong Wang and Peixi Peng. Open-r1-video. `https://github.com/Wang-Xiaodong1899/Open-R1-Video`, 2025. 1

[22] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 5

[23] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 4

[24] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 2, 4, 7

[25] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3

[26] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*, 2025. 5

[27] Xingjian Zhang, Xi Weng, Yihao Yue, Zhaoxin Fan, Wenjun Wu, and Lei Huang. Tinyllava-video: A simple framework of small-scale large multimodal models for video understanding. *arXiv preprint arXiv:2501.15513*, 2025. 1, 2, 3

[28] 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. 4

[29] 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. 6

[30] 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. 1, 2, 3

[31] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 6