Reinforcing Video Reasoning with Focused Thinking

Jisheng Dang[1,2,5†], Jingze Wu[2†], Teng Wang[3*], Xuanhui Lin[1], Nannan Zhu[2], Hongbo Chen[2], Wei-Shi Zheng[2], Meng Wang[4], and Tat-Seng Chua[5]

[1] Lanzhou University, Lanzhou, China
[2] Sun Yat-sen University, Guangzhou, China
[3] University of Hong Kong, Hong Kong, China
[4] Hefei University of Technology, Hefei, China
[5] National University of Singapore, Singapore

Abstract. Recent advancements in reinforcement learning, particularly through Group Relative Policy Optimization (GRPO), have significantly improved multimodal large language models for complex reasoning tasks. However, two critical limitations persist: 1) they often produce unfocused, verbose reasoning chains that obscure salient spatiotemporal cues, and 2) binary rewarding fails to account for partially correct answers, resulting in high reward variance and inefficient learning. In this paper, we propose TW-GRPO, a novel framework that enhances visual reasoning with focused thinking and dense reward granularity. Specifically, we employ a token weighting mechanism that prioritizes tokens with high informational density (estimated by intra-group information entropy), suppressing redundant tokens like generic reasoning prefixes. Furthermore, we reformulate RL training by shifting from single-choice to multi-choice QA tasks, where soft rewards enable finer-grained gradient estimation by distinguishing partial correctness. Additionally, we propose question-answer inversion, a data augmentation strategy to generate diverse multi-choice samples from existing benchmarks. Experiments demonstrate superior performance on several video reasoning and understanding benchmarks. Notably, under identical training settings, our method significantly outperforms the Video-R1, achieving an average improvement of 2.6% across four comparable benchmarks, e.g., +3.8% on MVBench, +2.4% on TempCompass, and +2.3% on VideoMME. Our codes are available at https://github.com/longmalongma/TW-GRPO.

1 Introduction

Recent advances in reinforcement learning (RL) for large language models (LLMs) have yielded significant improvements in reasoning capabilities. DeepSeek-R1 [13] demonstrated that pure RL optimization can substantially improve model reasoning, while subsequent works [8; 25; 27; 42; 52; 53] extended these benefits to

† Equal contribution. * Corresponding author.

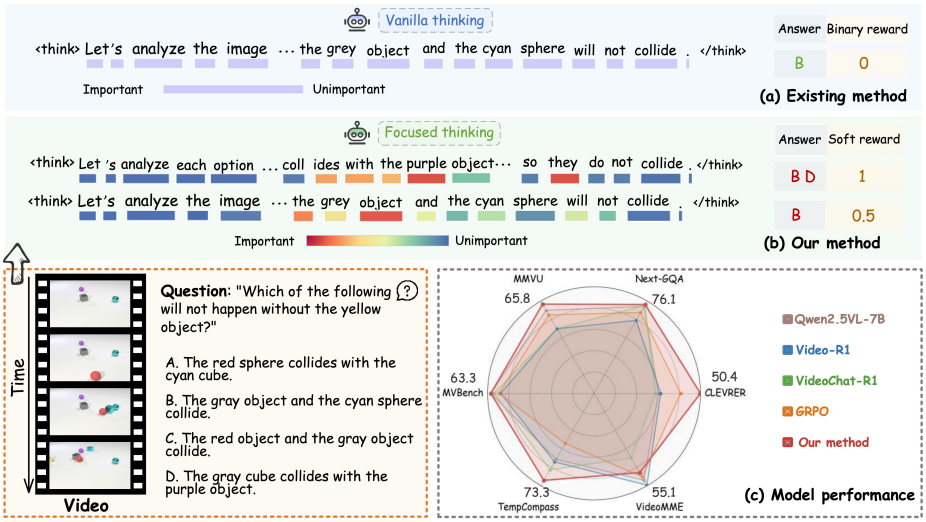

Fig. 1: TW-GRPO integrates focused thinking and soft multi-level rewards for multi-choice QA. Unlike vanilla thinking, which assigns uniform token importance, focused thinking highlights critical tokens to dominate loss calculation. By shifting single-choice QA's binary rewards to multi-choice QA's multi-level rewards, TW-GRPO enables fine-grained gradient estimation and training efficiency.

multimodal scenarios. Notable examples include Video-R1 [11], which introduced T-GRPO for spatiotemporal reasoning, VideoRFT [34] that proposes a semantic-consistency reward to reduce hallucination in reasoning, and VideoChat-R1 [17] that leverages GRPO-based multi-task joint fine-tuning. These improvements show promising progress in the understanding of fine-grained video details and multi-step reasoning.

Although RL-based approaches excel in optimizing verifiable metrics, critical challenges persist in refining reasoning quality and reward granularity for complex multimodal tasks. **First**, while chain-of-thought (CoT) reasoning has proven effective for solving language-based intricate problems [31], its application to MLLMs often results in verbose, unfocused thinking chains (*e.g.*, overthinking [14]). Building upon this, current training objectives fail to prioritize semantically critical spatio-temporal cues, which may obscure pivotal information and hurt learning efficiency. **Second**, existing methods rely on sparse, binary rewards derived from single-choice question-answer (QA) tasks [11, 25, 44]. These rewards assign maximum credit for fully correct answers and none otherwise, disregarding partial correctness. Recent work on video grounding [17], shows that soft reward signals enable finer-grained optimization. However, such approaches remain underexplored in mainstream video QA tasks, where single-choice formats lack naturally defined multi-level reward signals.

To address these limitations, we propose TW-GRPO, a novel framework that enhances GRPO via token weighting and multi-grained rewards. As shown in Fig. 1, our dynamic weighting mechanism prioritizes tokens with high informa-

tional density during loss computation. Specifically, we estimate token importance by analyzing intra-group information entropy across token positions, focusing the model on content critical to reasoning outcomes rather than generic phrases (*e.g.*, prefatory statements and repeated verifications). By prioritizing these tokens, the model learns to generate concise, task-aware reasoning chains while avoiding redundant or irrelevant details.

Furthermore, we reformulate RL training using multi-choice QA tasks, replacing sparse binary rewards with multi-level ones. Our approach distinguishes between partially correct and fully incorrect answers, enabling finer-grained gradient estimation and stabilizing policy updates. To mitigate multi-choice data scarcity, we introduce Question-Answer Inverse (QAI), which converts single-choice into multi-choice formats by negating questions and inverting answers.

Experiments demonstrate TW-GRPO's remarkable data efficiency and effectiveness, achieving state-of-the-art (SOTA) performance across multiple video benchmarks. With only 1K reinforcement learning samples, our model already achieves SOTA on five benchmarks, outperforming models like VideoChat-R1, which uses 18 times more data. When trained with 4K samples, our method surpasses the heavily pretrained Video-R1 on key benchmarks such as NExT-GQA and MVBench, with improvements of 2.0% and 1.5% respectively (Table 1). With focused thinking, qualitative analysis reveals condensed reasoning chains focused on critical visual or logical cues. Multi-level rewards also reduce reward variance during training. Our contributions are summarized as follows:

- We propose token weighting, a mechanism prioritizing tokens with high informational density during loss computation, enabling concise reasoning.
- We propose multi-grained reward modeling using multi-choice QA with partial correctness evaluation, improving gradient estimation and policy stability.
- We propose question-answer inverse, a data augmentation converting single-choice QA into multi-choice formats via question negation and answer inversion, mitigating data scarcity.

2 Related Works

2.1 Enhancing Reasoning in MLLMs

Enhancing LLM reasoning with reinforcement learning frameworks such as GRPO [31] has become an active research direction. In the multimodal setting, a growing body of work [6; 11; 17; 27; 33; 43; 44; 53] adopts GRPO-style optimization with verifiable rewards to improve visual reasoning. However, because GRPO typically ties supervision to final-answer correctness, it suffers from the uniform credit assignment problem [29]: all tokens in a reasoning chain receive the same reward, even though many tokens (e.g., generic scaffolding like "Let's think..."), as illustrated in Fig. 1) contribute little to task performance. This mismatch can produce inefficient learning and encourage verbose or redundant rationales.

To mitigate coarse token-level supervision, prior work has introduced external verifiers for image/video captioning [10; 24], or leveraged intrinsic signals for single-modal reasoning [6; 50]. Despite effectiveness, these methods can be computationally expensive and are often ill-suited for video reasoning, which requires extracting fine-grained spatio-temporal information. Recently, [47] showed that position-aligned distributional divergence across samples can serve as a robust indicator for detecting incorrect reasoning paths. Different from the line of work, we treat token importance as an explicit training signal rather than merely a post-hoc indicator, using it to modulate optimization at the token level and allocate stronger learning signals to more informative tokens. Moreover, to the best of our knowledge, we are the first to apply token-importance modeling to multimodal LLMs, enabling a token-level extension of GRPO that better matches multimodal reasoning dynamics and improves training efficiency.

2.2 MLLMs for Video Understanding

Video understanding is a crucial capability for MLLMs, enabling them to interpret and reason over dynamic visual content [7; 32; 38; 48; 49]. Recent advancements have resulted in the development of MLLMs specifically designed to improve video understanding tasks. Recent models such as VideoLLaMA2 [7] improve video-language alignment through spatiotemporal modeling and multimodal integration. Building on the success of DeepSeek-R1 [13], the GRPO algorithm has been widely adopted in video reasoning tasks, such as Video-R1 [11] for temporal reasoning and VideoRFT [34] for context-aligned reasoning. However, existing frameworks rely on sparse, binary reward signals [11; 34], offering no distinction between partially correct and entirely incorrect responses in video QA. Recently, VideoChat-R1 [17] introduced an IoU-based soft reward for video grounding, providing continuous feedback that has proven effective in improving learning stability and precision. But this approach is restricted to video grounding tasks and cannot be directly applied to the Video-QA task. In this work, we explore the use of soft reward in Video-QA. Specifically, we reformulate the RL objective as a multi-choice problem, enabling multi-level assignment and fine-grained optimization.

3 Methodology

3.1 Preliminary

As shown in Fig. 2, we focus on the task of multi-choice Video-QA, where the model is required to select the correct answer from a set of candidate options based on both the video content and the question. To enhance the performance of MLLMs on this task, GRPO [31] has been introduced. For an input query q, GRPO samples G candidate responses $o = \{o_1, \ldots, o_G\}$ from the policy model. Then, the rule-based reward assesses these responses to obtain reward scores

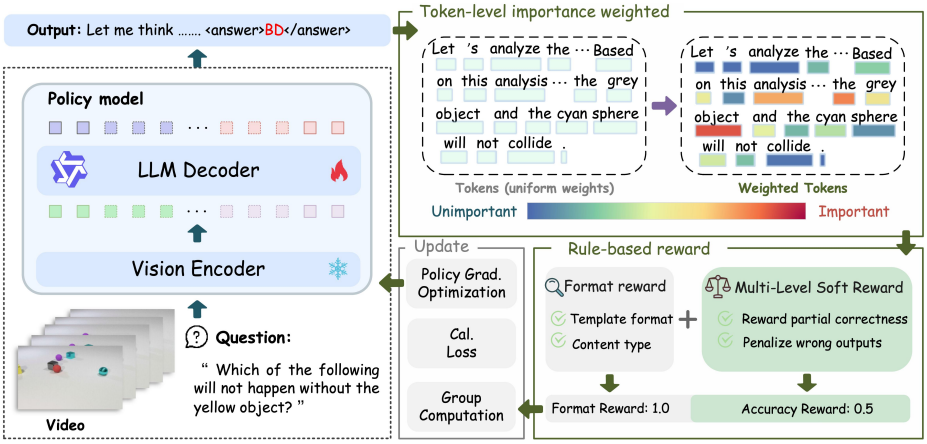

Fig. 2: Overview of the TW-GRPO framework. The diagram shows the key steps in a forward pass, starting from the video input, generating candidate completions, and computing the reward, with adjustments to the final objective and model updates. Specifically, a multi-level soft reward is incorporated into the reward calculation to provide partial correctness feedback. These signals are then integrated into the final objective, where token-level importance weighting is applied, allowing the model to prioritize more informative tokens and improve overall performance.

$\{R_1, \ldots, R_G\}$. The relative quality of o_i is then computed through standardization:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}, \tag{1}$$

where $\hat{A}_i$ denotes the normalized advantage of the i-th response within the group. The optimization objective combines response improvement with regularization:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = & \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\ & \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(r_{i,t}(\theta)\hat{A}_i,\right.\right.\right. \\ & \left.\left.\left.\text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\Big)\hat{A}_i\right) - \beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\right)\right], \end{aligned} \tag{2}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{3}$$

Despite its efficiency, existing GRPO algorithms do not differentiate between token positions during optimization, as shown in Equation 2. This leads the model to expend unnecessary effort on uninformative tokens, such as generic phrases like "Let's think...", which are less useful for policy optimization than tokens that describe critical spatiotemporal cues. In addition, since QA tasks are typically formulated as single-choice classification problems, existing approaches

often adopt binary reward signals. However, in video grounding tasks [17], such binary signals have been shown to be less effective than continuous reward. Nevertheless, how to incorporate multi-level rewards into single-choice QA tasks remains unexplored.

To address these two issues, we propose the **TW-GRPO** framework, as illustrated in Fig. 2. To address the overlooked variation in token importance, we introduce a *token-level importance weighting* mechanism in Section 3.2 that guides the model to focus on informative tokens. Furthermore, to overcome the limitations of binary rewards in single-choice QA, we reformulate the task as a *multi-answer* setting, where a question may have one or more correct options. Then, a crafted *multi-level soft reward* is designed to provide partial credit, enabling finer-grained policy learning, which is introduced in Section 3.3.

3.2 Token-Level Importance Weighting

In this section, we address how to model token-level importance, enabling the policy to distinguish informative tokens better and optimize more effectively. As usual, the fine-grained assessment of reasoning quality typically requires an auxiliary critic model [44], which introduces additional parameters and undermines one of GRPO's main advantages. Inspired by recent work [4; 19; 47], which demonstrates that key reasoning tokens can be identified based on token-level distributional differences, we propose a lightweight approach grounded in the concept of information entropy. The key insight is that token positions in which candidate outputs exhibit greater divergence from the expected distribution are more likely to carry more information. This allows us to estimate token importance without introducing extra model components.

Specifically, we propose the token importance weight w_t to quantify the information content of each token position. In detail, the Kullback-Leibler (KL) divergence D_{KL} measures the discrepancy between the probability distribution of the token at position t in the candidate output o_i and the expected distribution at the same position. For token position t, we calculate the divergence D_t as:

$$D_t = \sum_{i=1}^{G} D_{\mathrm{KL}} \left(p(o_{i,t}) \middle\| \mathbb{E}[o_t] \right), \tag{4}$$

where G denotes the number of candidate outputs, and $\mathbb{E}[o_t]$ is the expected probability distribution at token position t, which is computed by averaging the probability distributions of each candidate output, with the missing tokens filled in using the uniform distribution $\mathcal{U}(V)$ to account for variable sequence lengths [15]. This filling process ensures that all token sequences, regardless of length, contribute equally to the divergence calculation, preventing bias toward longer sequences and capturing meaningful information in shorter sequences. The core intuition behind this measurement is that high divergence at specific token positions reflects greater variability in model predictions. As an empirical observation, we find that such high-weight positions do not scatter randomly, but consistently concentrate on key reasoning tokens (*e.g.*, critical entities, actions,

and logical operators), as visualized in Fig. 5 and Fig. A4. Based on this observation, our optimization process dynamically allocates stronger learning signals to these positions, enhancing training efficiency. To ensure numerical stability and comparable importance scores across different positions, we normalize the divergence measurements using min-max normalization:

$$w_t = (1+\alpha)\cdot \frac{D_t - D_{\min}}{D_{\max} - D_{\min}}. \tag{5}$$

In this formulation, α is a hyperparameter that controls the scaling of token importance. To ensure comparability, the raw divergence scores are normalized to a standard range while preserving their relative differences. Moreover, the addition of the constant offset $(1+\alpha)$ guarantees that tokens with low divergence retain non-zero weights, thus preventing any position from being entirely ignored during training. Consequently, the resulting weights w_t enable position-sensitive optimization by modulating the learning signal according to token-level informativeness. Further details on our normalization and padding strategies are available in Appendix E.3.1 and Appendix E.3.2.

$$\mathcal{J}_{\text{TW-GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(w_t r_{i,t}(\theta)\hat{A}_i,\quad \text{clip}\big(r_{i,t}(\theta),1-\varepsilon,1+\varepsilon\big)\hat{A}_i\right)\right]. \tag{6}$$

Here, we normalize the number of outputs $|o_i|$ to balance their contributions to the overall loss [41]. Notably, our method does not require additional evaluation models to assess each step of the model's output. It only requires simple distance calculations to guide the model in applying varying levels of attention to different reasoning positions, and the computational complexity of this mechanism scales strictly linearly with the sequence length. After addressing the first issue, we focus on the second challenge: the need for multi-level reward signals in QA tasks.

3.3 Multi-Answer Soft Reward

In this section, we address the inefficiency of reward signals in QA tasks caused by the single-choice formulation. Our solution consists of two main steps. **First**, inspired by multi-choice formats in standardized testing, we reformulate the single-choice QA task as a *multi-answer* setting, in which each question contains at least one correct answer, possibly more. **Second**, leveraging this multi-choice question, we introduce a *multi-level soft reward* that assigns partial credit based on the overlap between the predicted and ground-truth answers. This enables more informative reward feedback for reinforcement learning. We describe each component in detail below.

Redefine Multi-choice QA Task. In Video-QA tasks, standard benchmarks such as NExT-GQA [37] are typically formulated as single-choice questions, where each question has exactly one correct answer. As a result, evaluation is based

on 0/1 accuracy, which inherently limits the ability to generate multi-level soft rewards. Interestingly, multi-answer questions, which are often used in the most challenging sections of standardized exams, align well with the needs. These questions offer a suitable level of difficulty and include at least one correct option, making them well-suited to our training objective. Therefore, we reformulate the single-choice QA task as a multi-answer one. Unfortunately, this raises a new challenge: obtaining suitable multi-answer data for training. To the best of our knowledge, only a limited number of existing datasets, such as the counterfactual reasoning task in CLEVRER [39], address multi-choice problems.

To mitigate this scarcity, we introduce question-answer inversion, a novel data augmentation technique that transforms single-choice questions into multi-answer questions in general datasets. For example, in the NExT-GQA [37] dataset, a question such as "Why did the boy pick up one present from the group and move to the sofa?" is modified by changing "did" to "didn't," thereby transforming it from a five-choice, single-answer question into a five-choice, multiple-answer question. To prevent the model from incorrectly associating negation with selecting multiple correct answers, we introduce a mechanism that randomly removes correct options, ensuring that the model remains challenged and is required to reason more carefully. Finally, by performing random question-answer inversion on the original dataset, we construct the final multi-choice NExT-GQA dataset, which may contain more than one correct answer, thereby increasing task complexity. To validate the quality of QAI, we evaluated 528 inverted samples from 1,000 NExT-QA instances using Qwen3-VL-235B-A22B-Instruct [2] as an LLM judge, supplemented by a 50-sample human review. We assessed **logical validity** (whether the new correct answers represent events confirmed absent under the negated premise) and **spurious artifacts** (whether grammatical shortcuts allow answer selection without video observation). The LLM judge yields an 86.0% validity rate and a 4.0% artifact rate, consistent with the 96.0% validity rate observed in the human review. These results confirm that QAI generates predominantly valid counterfactual training signals. Representative examples and the complete evaluation prompt are provided in Appendix B.

However, while introducing multi-choice questions improves the training environment for RL, it also presents significant challenges. The increased complexity causes traditional accuracy-based reward mechanisms, which rely on binary accuracy (0 or 1), to exhibit significant reward variance between single-choice and multi-choice questions. As shown in Fig. 3, the GRPO model trained on the single-choice dataset (GRPO(single-choice)) exhibits a notably lower reward standard deviation compared to the GRPO model trained on the multi-choice dataset (GRPO(fixed reward)). This higher variance complicates model convergence, making stable improvements in reasoning ability more difficult. Therefore, a key challenge is to optimize the fixed reward mechanism to mitigate this variance and ensure consistent progress in reasoning over the multi-choice dataset.

Multi-Level Soft Reward. To address this, we draw inspiration from the IoU reward used in grounding tasks such as VideoChat-R1 [17], where a continuous-valued reward reflects the degree of temporal overlap between predicted and

ground-truth intervals. This soft feedback enables the model to capture partial correctness and optimize both precision and recall, rather than relying solely on exact matches. We proposed the multi-level soft reward, which assigns graded credit to correct predictions partially and penalises incorrect ones, defined as:

$$R_{\text{soft}} = \begin{cases} \frac{|P|}{|G|}, & \text{if } P \subseteq G, \\ 0, & \text{if } P \not\subseteq G. \end{cases} \tag{7}$$

Here, P denotes the predicted set and G the ground truth set. If $P \subseteq G$, the reward is $|P|/|G|$; otherwise, it is 0. As shown in Fig. 1, if the ground truth is $\{B, D\}$ and the model predicts $\{B\}$, the reward is $1/2$, reflecting partial correctness. Predictions that include elements outside G receive a reward of 0, thereby penalizing false positives. This strict zero-penalty for false positives is chosen to prevent reward hacking, *e.g.*, a model selecting all available options to guarantee partial recall. Empirical comparison against softer alternatives (F1-score and no-penalty variants) confirms that this design consistently yields superior performance across benchmarks (Appendix E.3.6). In this way, multi-level soft reward ensures that the model is proportionally rewarded for partially correct answers, improving fine-grained gradient estimation and policy stability.

4 Experiment

4.1 Experiment Setup

We train our model on Qwen2.5-VL-7B using two NVIDIA H800 GPUs with a lightweight setup of 500 RL steps on 1,000 CLEVRER counterfactual training datasets, a subset of the Video-R1 dataset [11]. Under this setup, the total training time of TW-GRPO is only 4 hours, compared to 6 hours for standard GRPO. For a fair comparison with prior work, we also train a model on the broader Video-R1 dataset [11], using 1,000 RL steps on 4,000 samples as Video-R1-Zero [11]. Each frame is processed at a resolution of $128 \times 28 \times 28$ during training. For reasoning, the frame resolution is increased to $256 \times 28 \times 28$ to improve performance. Regarding computational overhead, QAI is performed offline via string matching before training and introduces no additional cost during the RL process. The token-level importance weighting operates with $\mathcal{O}(L)$ complexity with respect to the output sequence length, adding a negligible overhead of less than 0.01 seconds per training step in practice.

In particular, evaluation is conducted on six multi-choice QA video benchmarks covering general understanding and reasoning: MVBench [16], TempCompass [21], VideoMME [12], MMVU [51], NExT-GQA [37], and CLEVRER [39]. MVBench, TempCompass, and VideoMME emphasize general video understanding, combining visual perception and temporal comprehension without explicit reasoning focus. CLEVRER, NExT-GQA, and MMVU assess complex spatiotemporal and multimodal reasoning over dynamic videos. Together, these benchmarks comprehensively evaluate both general video understanding and fine-grained reasoning capabilities, ensuring an accurate assessment of the performance.

For all evaluations, we used the decoding configuration from the official Qwen2.5-VL demo, with top_p = 0.001 and temperature = 0.01. And we adopted the same experimental settings as in Video-R1 [11], including the prompt and revolution, and set the batch size to 16. For the NExT-GQA, MMVU, MVBench, TempCompass, and VideoMME datasets, we used the prompt words from Video-R1. Detailed settings are provided in Appendix D.

4.2 Main Results

Comparison with State-of-the-Arts. The results presented in Table 1 indicate that our proposed TW-GRPO achieves highly competitive results, including leading performance on several benchmarks, while requiring significantly fewer training samples than competing methods. On video reasoning tasks, our method demonstrates strong capabilities. It outperforms key reinforcement learning baselines on reasoning benchmarks such as NExT-GQA and MMVU, and achieves a high accuracy of 50.4% on CLEVRER. When trained on only 4K samples, it also achieves top performance on NExT-GQA, with 76.3% accuracy. These results suggest the effectiveness of our token-level weighting and soft reward mechanisms for complex reasoning. For general video understanding, TW-GRPO is highly competitive. It achieves the highest score among all compared methods on TempCompass at 73.3%. To further highlight its data efficiency, a direct comparison with Video-R1-Zero under an identical 4K training data setup shows that TW-GRPO outperforms it on all comparable benchmarks, including MVBench (+3.8%) and TempCompass (+2.4%). Crucially, a direct comparison under the 1K training condition also shows that TW-GRPO significantly improves upon GRPO across five of the six benchmarks. Importantly, because the Video-R1 dataset contains only open-ended and single-choice QA formats, the multi-level soft reward remains inactive during this phase. This naturally isolates the token-level importance weighted mechanism, demonstrating that it can achieve competitive performance on its own. These results underscore the data efficiency of our approach. Enhancing reinforcement learning with token-level importance weighting and multi-level reward strategies yields more stable policy learning, thereby boosting model performance across diverse tasks. Additional experiments are provided in Appendix E.

Training Dynamics and Convergence Behavior. Fig. 3 illustrates the training dynamics of different GRPO variants. Fig. 3(a) shows that TW-GRPO achieves faster convergence in reward standard deviation, indicating more stable learning. This stability is attributed to the introduction of multi-level soft reward and token-weighting strategies, which help the model handle ambiguous questions more effectively. In detail, traditional GRPO suffers from slow convergence on multi-choice tasks due to fixed accuracy rewards. In contrast, our soft reward strategy reduces the reward standard deviation, leading to more stable optimization. Furthermore, the token-level importance weighting mechanism, which helps the model focus on more informative tokens, thereby improving optimization efficiency and speeding up convergence. As shown in Fig. 3(b), TW-GRPO produces shorter

Table 1: Comparison of model performance on video reasoning benchmarks and general video understanding benchmarks. †represent training in Video-R1 data [11].

Video reasoning benchmarks				
Models	**Training**	**CLEVRER**	**Next-GQA**	**MMVU$_{_mc}$**
Baseline				
LLaMA-VID [18]	-	-	-	-
VideoLLaMA2 [7]	-	-	-	44.8
LongVA-7B [45]	-	-	-	-
Video-UTR-7B [40]	-	-	-	-
Kangaroo-8B [20]	-	-	-	-
Qwen2.5-VL-7B (Zero-Shot) [3]	-	30.5	75.9	65.4
Qwen2.5-VL-7B (CoT) [11]	-	27.7	73.4	63.0
Supervised Finetuning				
Qwen2.5-VL-7B (SFT) [11]	165K SFT	29.0	69.0	61.3
Reinforcement Learning Finetuned				
Video-R1-Zero [11]	4K RL	-	-	63.8
Video-R1 [11]	165K SFT + 4K RL	31.6	74.3	64.2
VideoRFT [34]	102K SFT + 8K RL	12.8	74.2	64.5
VideoChat-R1 [17]	18K RL	29.2	76.0	64.2
GRPO (Ours)	1K RL	41.1	75.2	65.1
TW-GRPO (Ours)	1K RL	**50.4**	<u>76.1</u>	**65.8**
TW-GRPO (Ours)†	4K RL	32.0	**76.3**	<u>65.6</u>

Video understanding benchmarks				
Models	**Training**	**MVBench**	**TempCompass**	**VideoMME$_{wo_sub}$**
Baseline				
LLaMA-VID [18]	-	41.9	45.6	-
VideoLLaMA2 [7]	-	54.6	-	47.9
LongVA-7B [45]	-	-	56.9	52.6
Video-UTR-7B [40]	-	58.8	59.7	52.6
Kangaroo-8B [20]	-	61.1	62.5	56.0
Qwen2.5-VL-7B (Zero-Shot) [3]	-	63.3	72.5	<u>56.5</u>
Qwen2.5-VL-7B (CoT) [11]	-	57.4	72.2	53.1
Supervised Finetuning				
Qwen2.5-VL-7B (SFT) [11]	165K SFT	59.4	69.2	52.8
Reinforcement Learning Finetuned				
Video-R1-Zero [11]	4K RL	60.4	70.9	53.8
Video-R1 [11]	165K SFT + 4K RL	62.7	72.6	**57.4**
VideoRFT [34]	102K SFT + 8K RL	61.4	72.1	54.1
VideoChat-R1 [17]	18K RL	63.1	<u>72.9</u>	52.4
GRPO (Ours)	1K RL	62.8	71.9	55.9
TW-GRPO (Ours)	1K RL	<u>63.3</u>	**73.3**	55.1
TW-GRPO (Ours)†	4K RL	**64.2**	**73.3**	56.1

output sequences, suggesting it learns to reason more concisely. This reflects a more substantial alignment between the reward objective and the final model behaviour, confirming the effectiveness of our proposed training design. As a result of this design, TW-GRPO achieves smoother convergence with fewer tokens in the generated outputs, reflecting more concise reasoning. Moreover, we provide an analysis of the dynamics of token weights in Appendices E.3.4 and E.3.5.

4.3 Ablation Study

To better understand the contributions of each component in our method, we conducted ablation studies to assess the improvements brought by indeterminate selection and TW-GRPO. In addition to using accuracy as an evaluation metric,

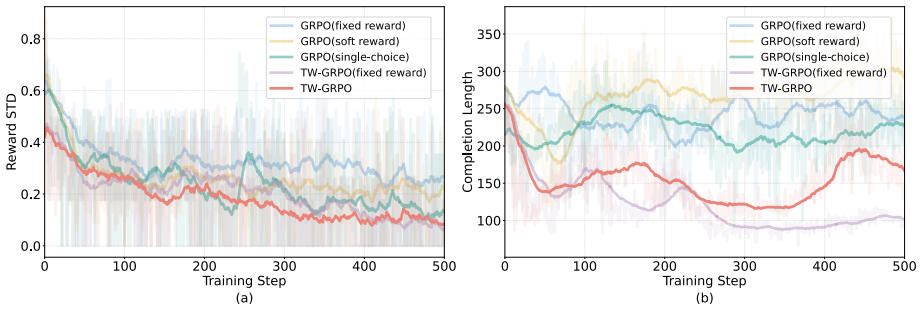

Fig. 3: Training dynamics of different GRPO variants. (a) TW-GRPO achieves faster convergence in reward standard deviation, indicating more stable learning. (b) It also outputs consistently shorter lengths, reflecting the reasoning effectively.

Table 2: Ablation study on data construction with different sources. We evaluate the effect of using single/multi-choice questions from CLEVRER, NExT-GQA, and STAR.

Setting	Single.	Multiple.		All		MMVU
	Acc.	Acc.	Soft Acc.	Acc.	Soft Acc.	Acc.
CLEVRER*$_{\text{cf}}$ *Source						
single-choice	**51.5**	18.2	50.9	32.6	51.2	62.7
multi-choice	50.4	**32.3**	**55.4**	**41.1**	**52.7**	**65.1**
NExT-GQA Source						
single-choice	48.0	**12.3**	**45.5**	30.7	45.7	64.3
multi-choice	**63.5**	9.6	44.9	**36.7**	**54.7**	**64.6**
STAR Source						
single-choice	65.1	**2.6**	**41.4**	**29.3**	**51.5**	64.8
multi-choice	**65.9**	1.5	40.8	29.1	**51.5**	**66.2**

we compute soft accuracy based on Equation 7, enabling a more fine-grained assessment of the model's performance on multi-choice datasets.

We evaluate the necessity of adding uncertain options by assessing the performance of the training model on datasets with different problem types. Specifically, the multi-choice training data used for the STAR and NExT-GQA datasets are generated through our question-answer inversion method. As shown in Table 2, for datasets such as CLEVRER, NExT-GQA, and STAR [36], multi-choice training outperforms training with only single-choice questions. The results demonstrate that incorporating multiple-choice questions increases the complexity of the training set, thereby enhancing the model's reasoning ability. For instance, the accuracy of the model trained on multi-choice questions in the CLEVRER dataset is 41.1%, which is 8.5% higher than that of the GRPO model trained on single-choice questions. Moreover, the counterfactual QA subset of CLEVRER [39] includes samples that involve object dynamics and hypothetical outcome predictions. This provides a more complex training environment.

Table 3: Ablation study on training type, reward design, and the token weighting. We evaluate performance on CLEVRER and MMVU under different settings.

Setting	Single.	Multiple.		All		MMVU
	Acc.	Acc.	Soft.	Acc.	Soft.	Acc.
Training Problem Type (TW-GRPO)						
single-choice	60.4	22.8	55.9	38.9	57.8	63.7
multi-choice	**60.9**	**42.5**	**64.4**	**50.4**	**62.9**	**65.8**
Reward Design (TW-GRPO, multi-choice)						
fixed reward	**64.9**	26.0	54.2	42.6	58.7	65.0
soft reward	60.9	**42.5**	**64.4**	**50.4**	**62.9**	**65.8**
Effect of Token Weighting (multi-choice, fixed reward)						
GRPO	50.4	**32.3**	55.4	41.1	52.7	**65.1**
TW-GRPO	**64.9**	26.0	**54.2**	**42.6**	**58.7**	65.0
Effect of Token Weighting (multi-choice, soft reward)						
GRPO	58.3	28.1	57.6	41.2	57.9	64.6
TW-GRPO($_{\alpha=0}$)	**64.7**	36.6	60.5	48.6	62.3	62.1
TW-GRPO	60.9	**42.5**	**64.4**	**50.4**	**62.9**	**65.8**

Next, we investigate the effectiveness of TW-GRPO across different training setups, reward strategies, and token weighting configurations, as summarized in Table 3. We first observe that training on multi-choice question types yields better performance than single-choice only settings, especially on CLEVRER, where multi-choice training improves accuracy by more than 10% compared to single-choice training. Regarding reward design, soft rewards consistently outperform fixed rewards across settings, improving soft accuracy by up to 7.8% in multi-choice tasks. This indicates that multi-answer soft reward provides more informative and stable supervision signals, especially for ambiguous or partially correct answers. Finally, we assess the effect of token weighting. In both fixed and soft reward settings, TW-GRPO shows notable gains over GRPO. For example, with soft reward, TW-GRPO improves multiple-choice soft accuracy from 57.6% to 64.4%. When token weights are removed (TW-GRPO) ($\alpha = 0$)), performance drops, confirming that token-level importance modeled is crucial for fine-grained optimization. Detailed analyses, including sensitivity analysis on the weighting hyperparameter α, are provided in Appendix E.

4.4 Qualitative Analysis

Qualitative Analysis of Reasoning Path. We compare Video-R1 and TW-GRPO on a physics-based density estimation task from the MMVU dataset. As shown in Fig. 4, a stone is first weighed in air (230 g) and then submerged in water, where its apparent weight drops to 138 g. Solving the task requires applying Archimedes' principle to derive the displaced volume from the buoyant force of 92 g and to compute the density as mass over volume.

The baseline Video-R1 model attempts this but incorrectly assumes a volume of 100 cm^3 ①, leading to a calculated density of 2.3 g/cm^3. It then mistakenly

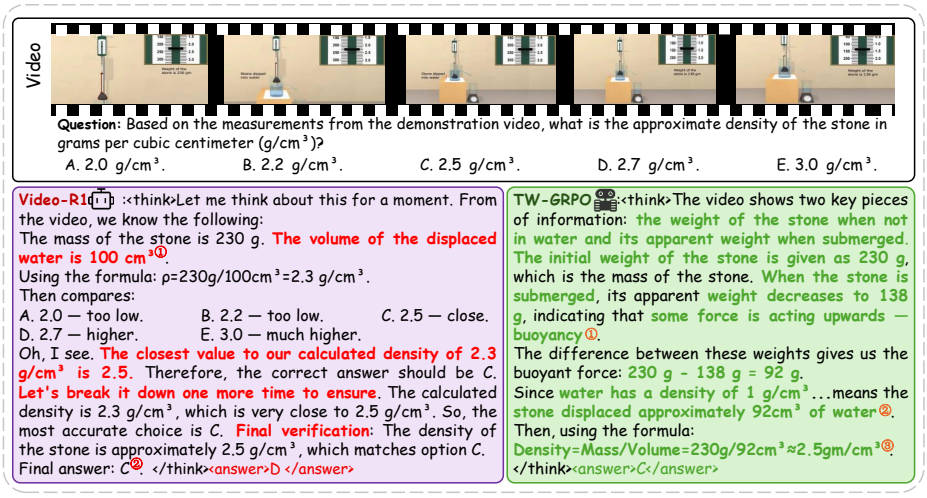

Fig. 4: Comparison of reasoning from Video-R1 and TW-GRPO on MMVU samples.

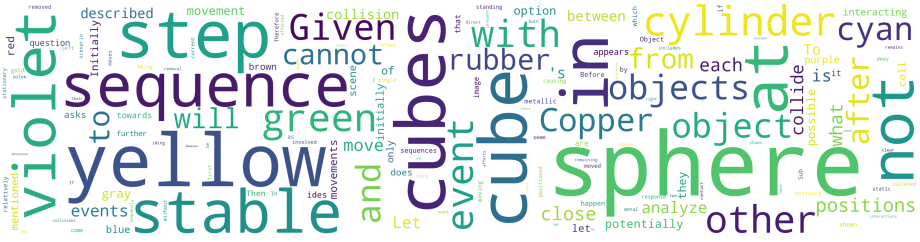

Fig. 5: Word cloud of focused tokens.

claims 2.5 g/cm^3 is the closest answer, despite 2.2 g/cm^3 being numerically closer ②. Although the model attempts to refine its performance through reflection, it remains anchored to an incorrect conclusion, resulting in substantial token inefficiency. Moreover, it ultimately selects a value of 2.7g/cm^3, contradicting its prior (and redundant) estimate. In contrast, TW-GRPO trained model accurately extracts the key values from the video, applies physical principles to infer volume from buoyant force ①. This stable logical chain leads to a standard density computation and precise option matching ②③. This case study clearly illustrates that TW-GRPO successfully suppresses hallucinated shortcuts, enforcing grounded, quantitative, and logically consistent reasoning from dynamic visual cues. Additional reasoning trajectories are detailed in Appendix F.

Qualitative Analysis of Focused Thinking. To investigate what the model prioritizes, we analyze the vocabulary generated at the highest-weighted sequence positions. Fig. 5 visualizes these terms as a word cloud. A distinct semantic pattern emerges. Specifically, the prominent vocabulary is densely populated with

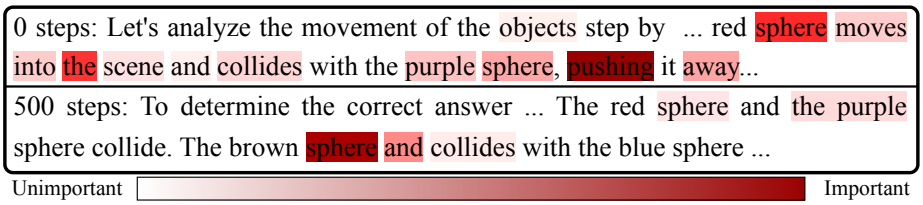

Fig. 6: Reasoning change after training.

concrete geometric entities (e.g., "sphere", "cylinder"), dynamic events (e.g., "collide", "stable"), and temporal markers (e.g., "sequence", "step"). This phenomenon strongly validates our core hypothesis. By optimizing solely on positional divergence, the mechanism naturally grounds the model's attention in the essential visual and logical cues required for complex reasoning.

Moreover, we observe a distinct progressive learning behavior in the reasoning chains. Fig. 6 tracks the positional weight distribution of a same sample from 0 to 500 training steps. Over time, the reasoning outputs transition from verbose text to highly dense structural logic. For instance, the description of a single collision event is reduced from 15 to 8 words. More importantly, the initially dispersed positional weights become sharply focused on critical spatiotemporal entities. This temporal evolution proves that TW-GRPO actively drives the model to eliminate reasoning redundancy and up-weight essential reasoning steps. Further visualizations are available in Appendix F.

5 Conclusions

In this work, we present TW-GRPO, a novel reinforcement learning framework that advances video reasoning in MLLMs. While prior approaches have improved model accuracy, two core limitations remain: the inability to distinguish token-level contributions and the inefficiency of binary reward signals. TW-GRPO addresses these challenges by incorporating token-level importance weighting and introducing multi-answer soft rewards that grant partial credit for partially correct responses. Extensive experiments across six benchmarks, along with comprehensive ablation studies, validate the effectiveness of our approach. With only 1K training samples, TW-GRPO achieves leading performance on five out of six benchmarks, demonstrating strong data efficiency. We hope this insight provides a foundation for future research in video reasoning with MLLMs.

Bibliography

[1] Agarwal, S., Zhang, Z., Yuan, L., Han, J., Peng, H.: The unreasonable effectiveness of entropy minimization in LLM reasoning. In: NeurIPS. vol. 38, pp. 107150–107180 (2025)

[2] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

[3] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)

[4] Bigelow, E., Holtzman, A., Tanaka, H., Ullman, T.: Forking paths in neural text generation. In: ICLR. vol. 2025, pp. 17134–17167 (2025)

[5] Chen, L., Li, L., Zhao, H., Song, Y., 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

[6] Chen, M., Chen, G., Wang, W., Yang, Y.: Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. arXiv preprint arXiv:2505.12346 (2025)

[7] Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)

[8] Dai, W., Chen, P., Ekbote, C., Liang, P.P.: Qoq-med: Building multimodal clinical foundation models with domain-aware GRPO training. In: NeurIPS. vol. 38, pp. 37406–37453 (2025)

[9] Du, W., Yang, Y., Welleck, S.: Optimizing temperature for language models with multi-sample inference. In: ICML. vol. 267, pp. 14648–14668. PMLR (2025)

[10] Fan, K., Feng, K., Lyu, H., Zhou, D., Yue, X.: Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. arXiv preprint arXiv:2505.17018 (2025)

[11] Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in MLLMs. In: NeurIPS. vol. 38, pp. 99114–99137 (2025)

[12] Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR. pp. 24108–24118 (2025)

[13] Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025)

[14] Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., Zhang, B., Fu, C., Gao, P., Li, H.: MME-CoT: Benchmarking

chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In: ICML. vol. 267, pp. 27793–27830. PMLR (2025)
[15] Kang, Z., Zhao, X., Song, D.: Scalable best-of-n selection for large language models via self-certainty. In: NeurIPS. vol. 38, pp. 19720–19745 (2025)
[16] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR. pp. 22195–22206 (2024)
[17] Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
[18] Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: ECCV. pp. 323–340. Springer (2024)
[19] Lin, Z., Liang, T., Xu, J., Liu, Q., Wang, X., Luo, R., Shi, C., Li, S., Yang, Y., Tu, Z.: Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. In: ICML. vol. 267, pp. 37906–37918. PMLR (2025)
[20] Liu, J., Wang, Y., Ma, H., Wu, X., Ma, X., Wei, X., Jiao, J., Wu, E., Hu, J.: Kangaroo: A powerful video-language model supporting long-context video input: J. liu et al. IJCV **134**(3), 114 (2026)
[21] Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., Hou, L.: Tempcompass: Do video llms really understand videos? In: ACL Findings. pp. 8731–8772 (2024)
[22] Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W.S., Lin, M.: Understanding r1-zero-like training: A critical perspective. In: 2nd AI for Math Workshop @ ICML 2025 (2025)
[23] Liu, Z., Yang, Z., Chen, Y., Lee, C., Shoeybi, M., Catanzaro, B., Ping, W.: Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy. arXiv preprint arXiv:2506.13284 (2025)
[24] Meng, D., Huang, R., Dai, Z., Li, X., Xu, Y., Zhang, J., Huang, Z., Zhang, M., Zhang, L., Liu, Y., et al.: Videocap-r1: Enhancing mllms for video captioning via structured thinking. arXiv preprint arXiv:2506.01725 (2025)
[25] Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., et al.: Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)
[26] Oh, B.D., Schuler, W.: Token-wise decomposition of autoregressive language model hidden states for analyzing model predictions. In: ACL. vol. 1, pp. 10105–10117 (2023)
[27] Peng, Y., Zhang, G., Zhang, M., You, Z., Liu, J., Zhu, Q., Yang, K., Xu, X., Geng, X., Yang, X.: Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536 (2025)
[28] Qin, T., Park, C.F., Kwun, M., Walsman, A., Malach, E., Anand, N., Tanaka, H., Alvarez-Melis, D.: Decomposing elements of problem solving: What" math" does rl teach? arXiv preprint arXiv:2505.22756 (2025)

[29] Samineni, S.R., Kalwar, D., Valmeekam, K., Stechly, K., Kambhampati, S.: RL in name only? analyzing the structural assumptions in RL post-training for LLMs. In: NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning (2025)
[30] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
[31] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
[32] Shu, F., Zhang, L., Jiang, H., Xie, C.: Audio-visual llm for video understanding. In: ICCV. pp. 4246–4255 (2025)
[33] Tian, X., Zou, S., Yang, Z., He, M., Waschkowski, F., Wesemann, L., Tu, P.H., Zhang, J.: More thought, less accuracy? on the dual nature of reasoning in vision-language models. In: ICLR (2026)
[34] Wang, Q., Yu, Y., Yuan, Y., Mao, R., Zhou, T.: VideoRFT: Incentivizing video reasoning capability in MLLMs via reinforced fine-tuning. In: NeurIPS. vol. 38, pp. 4350–4376 (2025)
[35] von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., Gallouédec, Q.: Trl: Transformer reinforcement learning. https://github.com/huggingface/trl (2020)
[36] Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: Star: A benchmark for situated reasoning in real-world videos. arXiv preprint arXiv:2405.09711 (2024)
[37] Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: CVPR. pp. 13204–13214 (2024)
[38] Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. In: ECCV. pp. 98–115. Springer (2024)
[39] Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. In: ICLR (2020)
[40] Yu, E., Lin, K., Zhao, L., Wei, Y., Zhu, Z., Wei, H., Sun, J., Ge, Z., Zhang, X., Wang, J., et al.: Unhackable temporal rewarding for scalable video mllms. arXiv preprint arXiv:2502.12081 (2025)
[41] Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. In: NeurIPS. vol. 38, pp. 113222–113244 (2025)
[42] Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., Huang, G.: Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In: NeurIPS. vol. 38, pp. 57654–57689 (2025)
[43] Zhan, Y., Zhu, Y., Zhao, H., Yang, F., Zheng, S., Tang, M., Wang, J.: Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. In: CVPR. pp. 5807–5817 (2026)
[44] Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In: ICCV. pp. 1859–1869 (2025)

[45] Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. TMLR (2025)
[46] Zhang, Q., Wu, H., Zhang, C., Zhao, P., Bian, Y.: Right question is already half the answer: Fully unsupervised LLM reasoning incentivization. In: NeurIPS. vol. 38, pp. 67345–67372 (2025)
[47] Zhang, T., Shi, H., Wang, Y., Wang, H., He, X., Li, Z., Chen, H., Han, L., Xu, K., Zhang, H., Metaxas, D.N., Wang, H.: TokUR: Token-level uncertainty estimation for large language mode reasoning. In: First Workshop on Foundations of Reasoning in Language Models (2025)
[48] Zhang, X., Peng, D., Zhang, Y., Guo, Z., Wu, C., Chen, C., Ke, W., Meng, H., Sun, M.: Towards self-improving systematic cognition for next-generation foundation mllms. arXiv preprint arXiv:2503.12303 (2025)
[49] Zhang, Y., Liu, Y., Guo, Z., Zhang, Y., Yang, X., Chen, C., Song, J., Zheng, B., Yao, Y., Liu, Z., et al.: Llava-uhd v2: an mllm integrating high-resolution feature pyramid via hierarchical window transformer. arXiv preprint arXiv:2412.13871 (2024)
[50] Zhao, X., Kang, Z., Feng, A., Levine, S., Song, D.: Learning to reason without external rewards. In: 2nd AI for Math Workshop @ ICML 2025 (2025)
[51] Zhao, Y., Zhang, H., Xie, L., Hu, T., Gan, G., Long, Y., Hu, Z., Chen, W., Li, C., Xu, Z., et al.: Mmvu: Measuring expert-level multi-discipline video understanding. In: CVPR. pp. 8475–8489 (2025)
[52] Zheng, R., Dou, S., Gao, S., Hua, Y., Shen, W., Wang, B., Liu, Y., Jin, S., Liu, Q., Zhou, Y., et al.: Secrets of rlhf in large language models part i: Ppo. arXiv preprint arXiv:2307.04964 (2023)
[53] Zhou, H., Li, X., Wang, R., Cheng, M., Zhou, T., Hsieh, C.J.: R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. arXiv preprint arXiv:2503.05132 (2025)