

# GRPO-CARE: CONSISTENCY-AWARE REINFORCEMENT LEARNING FOR MULTIMODAL REASONING

Anonymous authors

Paper under double-blind review

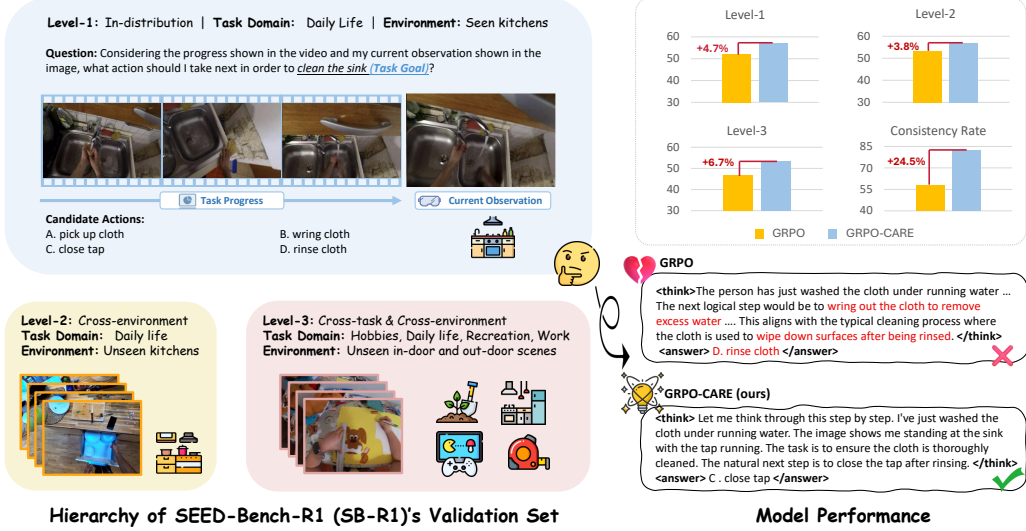


Figure 1: (a) **SEED-Bench-R1** (SB-R1) provides a systematic, three-level evaluation of post-training methods for MLLMs in video understanding, encompassing tasks that require both perception and reasoning to tackle complex real-world scenarios. (b) Our analysis identifies a key limitation of standard outcome-supervised GRPO: while it improves answer accuracy, it often compromises logical consistency between reasoning and answers. By introducing an adaptive, group-relative consistency bonus via reference-likelihood calibration, our **GRPO-CARE** achieves higher answer accuracy across all difficulty levels and improves interpretability, as reflected by increased consistency rates.

## ABSTRACT

Recent reinforcement learning (RL) approaches, such as outcome-supervised GRPO, have advanced Chain-of-Thought reasoning in large language models (LLMs), yet their adaptation to multimodal LLMs (MLLMs) remains underexplored. Progress has been further limited by the lack of evaluation settings that jointly test perception and reasoning under controlled generalization challenges. To enable such analysis, we reorganize prior benchmarks featuring complex real-world videos that demand intricate visual understanding and commonsense planning into **SEED-Bench-R1**, a structured testbed with large-scale training data and hierarchical evaluation across in-distribution, cross-environment, and cross-environment-task scenarios. Using this setting, we conduct a systematic experimental analysis of post-training methods, which reveals a key limitation of outcome-supervised GRPO: while it improves answer accuracy, it often compromises the logical coherence between reasoning and final answers, yielding only a 57.9% consistency rate. This stems from optimizing exclusively for final-answer rewards, which encourages shortcuts, and from rigid KL divergence penalties, which overly constrain adaptive reasoning. To address these issues, we propose **GRPO-CARE**, a novel consistency-aware RL framework that jointly optimizes correctness and coherence without requiring explicit process supervision. GRPO-CARE introduces a two-tiered reward: (1) a base reward for accuracy, and (2) an adaptive consistency bonus derived from a slowly evolving reference model that calibrates reasoning-to-answer likelihoods within peer groups. This mechanism rewards reasoning paths that are both correct and logically consistent, while removing the constraints of KL

penalties. Experiments on SEED-Bench-R1 show that GRPO-CARE consistently outperforms standard GRPO, achieving a 6.7% gain on the hardest evaluation level and a 24.5% increase in reasoning consistency. Moreover, models trained with GRPO-CARE transfer effectively to diverse video understanding and even language-only reasoning benchmarks, highlighting its robustness and generality.

## 1 INTRODUCTION

Recent progress in Large Language Models (LLMs) (Guo et al., 2025; Ope, 2024; Team et al., 2025) has been driven by advances in long Chain of Thought (CoT) generation, with reinforcement learning (RL) (Shao et al., 2024; Ouyang et al., 2022; Yeo et al., 2025) emerging as an effective post-training technique that improves complex problem-solving and generalization. Multimodal LLMs (MLLMs) extend these capabilities to process multimodal inputs (Zhang et al., 2025; Liu et al., 2025b; Meng et al., 2025), inheriting strong reasoning abilities while tackling richer, more complex data. However, current evaluations of RL-like post-training for MLLMs are fragmented: some focus narrowly on perception (e.g., detection, grounding) (Liu et al., 2025b), others on logical reasoning (e.g., multimodal math) (Huang et al., 2025), or else rely on broad datasets without structured generalization assessment (Feng et al., 2025).

We argue that *studying post-training for MLLMs requires evaluations that balance perception and logical reasoning, while rigorously testing generalization*. To this end, we reorganize prior benchmarks (Chen et al., 2023; Qiu et al., 2024) featuring complex real-world videos that demand intricate visual understanding and commonsense planning, and construct **SEED-Bench-R1** as a structured testbed. As shown in Fig. 1, SEED-Bench-R1 requires models to comprehend open-form task goals, track long-horizon visual progress, perceive complex environmental cues, and reason about next actions using world knowledge. It features hierarchical evaluation across three levels—(1) in-distribution, (2) OOD cross-environment, and (3) OOD cross-environment-task—with large-scale training data and verifiable ground-truth answers, making it suitable for supporting research in RL-like post-training methods for MLLMs.

Using SEED-Bench-R1, we conduct a comprehensive analysis comparing representative post-training approaches. Our experiments confirm that RL—specifically GRPO with outcome supervision (Shao et al., 2024)—is highly data-efficient and significantly outperforms supervised fine-tuning (SFT) on both in-distribution and OOD questions. However, we identify a key limitation: while outcome-supervised GRPO improves perception and answer accuracy for MLLMs, it often sacrifices logical coherence between reasoning chains and final answers, with a consistency rate of only 57.9%. This restricts interpretability and limits the potential performance ceiling. It originates from the fact that optimizing solely for final answer rewards creates a shortcut, where models prioritize correctness over maintaining logical reasoning, while strict KL penalties hinder adaptive adjustment of causal relations between reasoning and answers.

To overcome this, we propose **GRPO-CARE**, a novel RL framework with **Consistency-Aware Reward Enhancement** that jointly optimizes answer correctness and logical consistency without relying on explicit process supervision. As illustrated in Fig. 3, in addition to the base reward for answer correctness, we introduce a consistency bonus derived from a slowly-updated reference model through *likelihood calibration*. This bonus incentivizes the model to produce reasoning traces that are not only accurate but also logically coherent with the final answer. Specifically, GRPO-CARE maintains a reference model updated via exponential moving average (EMA) of the online model’s parameters, calibrating reasoning-to-answer consistency likelihoods. Samples that achieve both high accuracy and strong consistency are rewarded with an adaptive group-relative bonus, replacing rigid KL penalties and enabling more effective exploration of coherent reasoning paths.

Extensive evaluation shows that GRPO-CARE consistently outperforms standard GRPO across all difficulty levels, especially in challenging OOD scenarios, improving performance by 6.7% on Level-3 and increasing reasoning consistency by 24.5%. Ablation studies confirm that the consistency-aware reward is critical for balancing overall performance and interpretability, while transfer experiments to diverse video understanding benchmarks and even purely language-based reasoning tasks further demonstrate robustness and generality.

In summary, our main contributions are:

- A systematic reorganization of prior benchmarks into **SEED-Bench-R1**, providing a hierarchical and rigorous evaluation setting for studying post-training methods in multimodal reasoning.
- A comprehensive experimental study of post-training methods for MLLMs, revealing the limitations of outcome-supervised RL in maintaining logical coherence.
- A novel RL framework, **GRPO-CARE**, that introduces consistency-aware rewards to significantly improve reasoning interpretability and overall performance without explicit process supervision.

## 2 RELATED WORK

**RL for LLMs/MLLMs.** RL from human feedback (RLHF) aligns LLM outputs with human preferences via reward models trained on human preference data (Ouyang et al., 2022; Schulman et al., 2017). To enhance complex reasoning, generating long CoT is effective (Guo et al., 2025; Team et al., 2025; ope, 2024). RL methods like GRPO (Shao et al., 2024) and its variants DAPO (Yu et al., 2025b) and Dr.GRPO (Liu et al., 2025a) optimize CoT generation using outcome-based rewards. However, outcome-only supervision can yield inconsistent reasoning despite correct answers. Addressing this, some works train additional process supervision reward models with costly step-wise annotations (Lightman et al., 2023; Uesato et al., 2022; Chen et al.; Luo et al., 2024; Wang et al., 2023), incorporate LLM judges (Gao et al., 2024; Xia et al., 2025; Zhang et al., 2024a), or adaptive regularization via EMA-updated reference models (Ramé et al., 2024). In MLLMs, outcome-based RL may cause “Thought Collapse,” mitigated by stronger correctors (Wei et al., 2025) or step-wise reward matching (Zhang et al., 2025). Our GRPO-CARE employs a slowly updated reference model to provide bonus feedback on logically consistent and accurate responses, improving reasoning and accuracy without extra annotations or stronger correctors.

**Benchmarks for MLLM Post-training.** Recent RL-based post-training methods for MLLMs have primarily targeted image tasks—from perception (e.g., classification) to reasoning (e.g., visual math) (Huang et al., 2025; Liu et al., 2025b; Zhang et al., 2025; Sun et al., 2024). In contrast, video understanding, a more complex and general scenario, remains underexplored. Early RL-based efforts on video benchmarks (Wang & Peng, 2025; Zhao et al., 2025a) are limited by narrow tasks (e.g., emotion recognition) (Liu et al., 2022; Jiang et al., 2020) or scarce training data (Wu et al., 2024), hindering scalable analysis. Existing benchmarks (Li et al., 2024b; Liu et al., 2024b; Fang et al., 2024) mostly evaluate models post-trained on diverse general-domain data (e.g., Video-R1 (Feng et al., 2025)) but lack rigorous generalization assessment. To date, no comprehensive benchmark provides (1) large-scale training data for robust post-training, (2) structured validation sets across multiple generalization levels, and (3) multimodal questions balancing perception and reasoning in real-world scenarios. To address this, we propose SEED-Bench-R1, a video understanding benchmark with large-scale training data and a validation set partitioned into three generalization tiers, enabling comprehensive evaluation of MLLM post-training methods.

## 3 PILOT STUDY WITH SEED-BENCH-R1

### 3.1 SEED-BENCH-R1

**Benchmark Overview.** As shown in Fig. 1, SEED-Bench-R1 is built to systematically study how post-training methods affect multimodal reasoning in MLLMs. Building on EgoPlan-Bench (Chen et al., 2023) and EgoPlan-Bench2 (Qiu et al., 2024), it features: 1) real-world egocentric visual inputs, 2) diverse questions requiring commonsense reasoning for practical tasks, 3) a hierarchical validation design to assess robustness and generalization, and 4) large-scale, automatically constructed training questions with verifiable ground-truth answers.

**Visual Inputs and Question Design.** SEED-Bench-R1 leverages realistic egocentric videos (Damen et al., 2022; Grauman et al., 2022) of daily activities. Correctly answering its questions requires models to understand open-ended goals, track long-horizon task progress, perceive real-time environment states from an egocentric view, and apply world knowledge to infer the next action. The ground-truth answer comes from the actual next action occurring right after the current observation in the original uncropped video, with the negative options sampled from the same video. This challenging setting of candidate options demands a deep understanding of the environment state from dynamic visual input and world knowledge, such as action dependencies, rather than just semantic matching, to discern the correct action plan. Moreover, the derivation of golden answers is traceable and easy to verify.

Table 1: Statistics of SEED-Bench-R1, including a training set and a hierarchical three-level validation set for in-distribution, cross-environment, and cross-environment-task evaluations.

| Split  | # Samples | Domain                                | Cross-Env. | Cross-Task | Video Source  | Benchmark Source |
|--------|-----------|---------------------------------------|------------|------------|---------------|------------------|
| Train  | 50,269    | Daily life                            | -          | -          | Epic-Kitchens | EgoPlan-Bench    |
| Val-L1 | 2,432     | Daily life                            | ×          | ×          | Epic-Kitchens | EgoPlan-Bench    |
| Val-L2 | 923       | Daily life                            | ✓          | ×          | Ego4D         | EgoPlan-Bench    |
| Val-L3 | 1,321     | Hobbies, Daily life, Recreation, Work | ✓          | ✓          | Ego4D         | EgoPlan-Bench2   |

**Dataset Composition and Validation Levels.** As shown in Tab. 1, SEED-Bench-R1 includes both training and validation sets. The training set is automatically generated from Epic-Kitchens (Damen et al., 2022), covering daily household tasks. The validation set is human-verified and divided into three levels: **L1**: In-distribution evaluation with the same source and domain as training. **L2**: Cross-environment evaluation, using unseen kitchen environments from Ego4D (Grauman et al., 2022). **L3**: Cross-environment-task evaluation with the full Ego4D set, spanning hobbies, recreation, work, and daily life in diverse indoor and outdoor contexts, thus testing broader generalization.

### 3.2 EXPERIMENT SETUP

We use Qwen2.5-VL-Instruct-7B (Bai et al., 2025) as the backbone to study post-training on SEED-Bench-R1. Additional evaluation on different base models can be found in Appendix B.2. We adopt outcome-supervised GRPO (Shao et al., 2024) as a representative RL method and compare it with supervised finetuning (SFT). Each video is down-sampled to 16 frames at resolution  $128 \times 28 \times 28$ , plus one frame indicating the current observation. For SFT, training data is augmented with CoT reasoning distilled from Qwen2.5-VL-Instruct-72B and 7B via rejection sampling. GRPO instead relies on rule-based rewards without explicit CoT annotations. Following Guo et al. (2025), reasoning and answers are formatted within `<think>` `</think>` and `<answer>` `</answer>` tags.

Given a multimodal question  $x \sim \mathcal{D}$ , GRPO samples  $G$  responses  $\{o_g = (\tau_g, a_g)\}_{g=1}^G$  from the policy  $\pi_{\theta_{\text{old}}}$ , where  $\tau_g$  and  $a_g$  denote reasoning and answer. Unlike SFT, GRPO does not rely on predefined responses. The policy is optimized by maximizing:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{o_g\}} \frac{1}{G} \sum_{g=1}^G \frac{1}{|o_g|} \sum_{i=1}^{|o_g|} \min \left[ \frac{\pi_{\theta}(o_{g,i}|x, o_{g,<i})}{\pi_{\theta_{\text{old}}}(o_{g,i}|x, o_{g,<i})} \hat{A}_{g,i}, \right. \\ \left. \text{clip} \left( \frac{\pi_{\theta}(o_{g,i}|x, o_{g,<i})}{\pi_{\theta_{\text{old}}}(o_{g,i}|x, o_{g,<i})}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_{g,i} \right] - \beta \mathbb{D}_{KL}[\pi_{\theta} || \pi_{\text{ref}}]$$

Here,  $\varepsilon$  and  $\beta$  are hyperparameters, and  $\mathbb{D}_{KL}$  is the KL divergence between policy  $\pi_{\theta}$  and reference  $\pi_{\text{ref}}$ . The per-token advantage is set to the normalized reward  $\hat{A}_{g,i} = \tilde{r}_g = \frac{r_g - \text{mean}(\{r_1, \dots, r_G\})}{\text{std}(\{r_1, \dots, r_G\})}$ , with  $r_g$  computed by rules (e.g.,  $r_g = 1$  if the extracted answer matches ground truth, else 0).

### 3.3 RESULT ANALYSIS

Tab. 2 summarizes the performance of MLLMs trained with various methods on SEED-Bench-R1. Notably, compared to SFT, reinforcement learning with GRPO significantly improves data efficiency and boosts MLLM performance on both in-distribution (L1) and OOD (L2, L3) questions, despite relying only on a simple outcome-based reward without specialized CoT annotations.

Our analysis shows that GRPO mainly enhances perceptual abilities rather than reasoning. As shown in Fig. 2, the SFT-trained model is more prone to perceptual hallucinations, such as describing “a ball being hit from a tee” when this event does not occur. Attention map analysis reveals that GRPO-trained models generate CoT tokens that act as dynamic queries, attending to visual content more thoroughly—especially in OOD scenarios. For example, the GRPO model better highlights key visual observations and allocates more attention to critical objects (e.g., the ball on the tee), even if these are not explicitly referenced in the reasoning. We hypothesize that RL methods like GRPO encourage broader visual exploration via CoT, while SFT tends to produce superficial, pattern-memorized CoT with limited visual grounding. This likely underpins GRPO’s superior generalization.

However, outcome-supervised GRPO training for MLLMs has key limitations: unlike LLMs, MLLM reasoning does not improve proportionally during RL, often resulting in logical inconsistencies.

Table 2: Performance comparison on SEED-Bench-R1’s hierarchical validation set.

| Models           | L1 (In-Distribution) | L2 (Cross-Env) | L3 (Cross-Task, Cross-Env) |             |             |             | Overall     |
|------------------|----------------------|----------------|----------------------------|-------------|-------------|-------------|-------------|
|                  | Daily Life           | Daily Life     | Daily Life                 | Hobbies     | Recreation  | Work        |             |
| Qwen2.5-VL-7B    | 38.4                 | 40.1           | 35.8                       | 31.2        | 26.8        | 28.5        | 31.3        |
| SFT              | 46.2                 | 46.3           | 46.7                       | 41.7        | 44.3        | 38.4        | 42.7        |
| GRPO             | 52.3                 | 53.2           | 51.9                       | 43.7        | 55.2        | 39.4        | 46.7        |
| GRPO-CARE (ours) | <b>57.0</b>          | <b>57.0</b>    | <b>57.6</b>                | <b>51.2</b> | <b>57.4</b> | <b>48.5</b> | <b>53.4</b> |

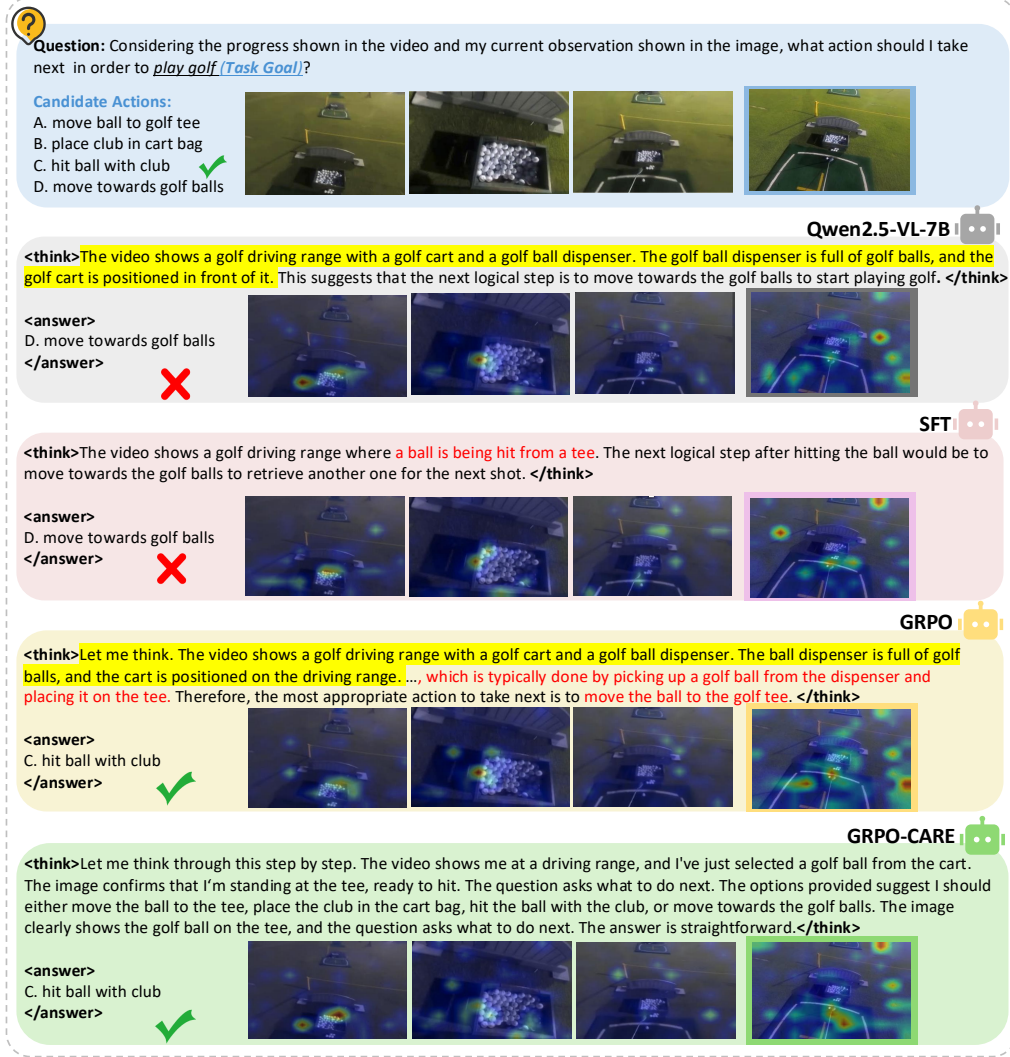


Figure 2: Case study of an L3 question from SEED-Bench-R1, showing a video of task progress, a final observation image, and attention maps (output-to-visual tokens). The **SFT** model tends to memorize reasoning patterns and exhibits perceptual hallucinations. The **GRPO** model attends more comprehensively to the highlighted key visual observation while lacking logical consistency in the generated content. The **GRPO-CARE** model further balances visual perception and logical reasoning.

While the GRPO-trained model frequently reaches correct answers, its CoT reasoning often lacks coherence. For instance, as shown in Fig. 2, initial reasoning steps mirror those of the base model (Qwen2.5-VL-7B), but later steps diverge and may contradict each other—e.g., suggesting “move the ball to the golf tee” but ultimately answering “hit ball with club.” Such inconsistencies, though sometimes yielding correct answers, undermine transparency.

Limited reasoning also constrains overall performance, as effective reasoning is crucial for integrating world knowledge with perception. For example, in Fig. 1, the GRPO model correctly identifies

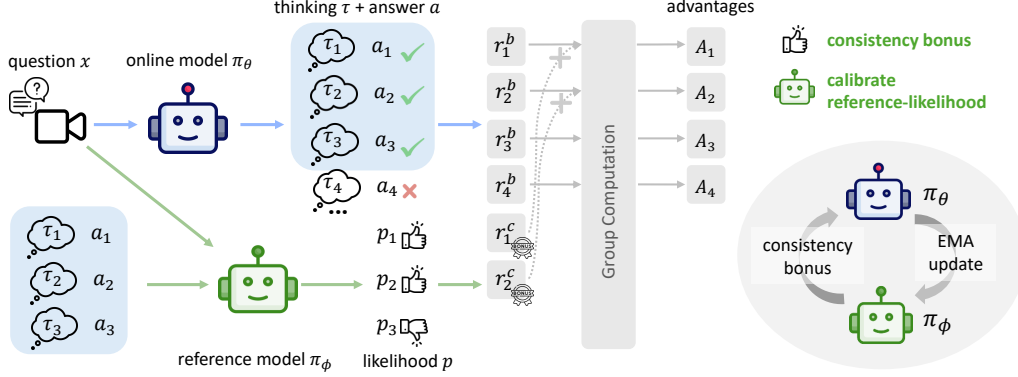


Figure 3: GRPO-CARE uses a two-tier reward system: a base reward for answer correctness ( $r_*^b$ ) and an adaptive consistency bonus ( $r_*^c$ ). The consistency bonus is given to high-accuracy samples whose reasoning-to-answer likelihood—estimated by a slowly updated (EMA) reference model—is higher than that of their group peers, conditioned on the multimodal question. The total reward, the sum of base and consistency rewards, is then used to compute advantages for updating the online model.

“running water” but fails to infer that the next logical step after cleaning is “turning off the faucet.” These reasoning-answer mismatches further complicate interpretability.

#### 4 CONSISTENCY-AWARE REWARD-ENHANCED GRPO (GRPO-CARE)

While outcome-supervised GRPO enhances visual perception in MLLMs, our analysis on SEED-Bench-R1 uncovers a critical trade-off: it often produces less logically coherent reasoning chains, thereby limiting interpretability and performance. This issue arises from two main limitations. First, the standard reward focuses exclusively on final-answer accuracy, overlooking the quality of intermediate reasoning steps. This can incentivize shortcut solutions—correct answers reached via inconsistent reasoning. Second, the KL penalty disproportionately constrains reasoning traces, typically longer than answers, thereby stifling exploration of diverse and coherent reasoning paths.

To address these challenges, we propose **GRPO-CARE** (Consistency-Aware Reward Enhancement), a method that jointly optimizes for both answer correctness and logical consistency, without requiring explicit supervision of the reasoning process. As shown in Fig. 3, GRPO-CARE introduces a two-tiered reward system: a base reward for answer correctness, and an adaptive consistency bonus. The consistency bonus is calculated by comparing the likelihood that a reasoning trace leads to the correct answer, as estimated by a slowly evolving reference model. For each high-accuracy sample generated by the online model, this likelihood is compared with those of its peers within the same group, encouraging the exploration of reasoning traces that are logically consistent with correct answers.

The training process, detailed in Algorithm 1, involves **two-stage filtering**. (1) First, we generate multiple reasoning traces per input and retain only those that exceed an accuracy baseline. (2) For these high-accuracy candidates, we assess how well each reasoning trace supports the final answer by calibrating its likelihood using a slowly evolving reference model.

**Reference Model and Likelihood Calibration.** The key insight is that a *stable reference model*—when conditioned on the online model’s reasoning trace—should assign a higher likelihood to the correct answer if the reasoning is logically grounded in the multimodal input. Specifically, the reference model is initialized from the same pretrained weights as the online model and updated smoothly via exponential moving average (EMA) to ensure stable likelihood estimation and self-adaptation. This setup encourages gradual self-evolution from a weaker to a stronger model without relying on a stronger external teacher, unlike methods that depend on fixed, stronger LLM judges. To avoid reinforcing “consistent-but-wrong” reasoning, we compute this likelihood only for trajectories with correct answers. Additionally, we cap the likelihood at a maximum threshold to prevent over-optimization toward artificially high values.

**Consistency Bonus Calculation.** Based on the clipped reference likelihoods, we compute a *group-relative consistency baseline* as the mean clipped likelihood (minus a small margin to avoid penalizing

**Algorithm 1** Consistency-Aware Reward Enhanced GRPO**Require:**


---

```

324  $\pi_\theta$ : Online policy model (initialized from pretrained weights)
325  $\pi_\phi$ : Reference model with EMA updates ( $\phi \leftarrow \theta$  initially)
326  $\mathcal{D}$ : Multimodal training dataset  $\{(x, y^*)\}$ 
327  $\lambda_{\text{cons}}$ : Consistency reward coefficient (e.g., 0.5)
328  $\gamma_{\text{acc}}$ : Minimum accuracy threshold (e.g. 0.1)
329  $\gamma_p$ : Maximum likelihood threshold (e.g. 0.95)
330  $\epsilon_p$ : Consistency margin (e.g. 0.01)
331
332 1: procedure TRAINING( $\pi_\theta, \mathcal{D}, T$ )
333 2:   for  $t \leftarrow 1$  to  $T$  do
334 3:     for each multimodal input  $x$  in batch  $\mathcal{D}$  do
335 4:       Phase 1: Trajectory Generation & Reward Computation
336 5:       Generate  $G$  reasoning traces + answers:  $\{\tau_g, a_g\}_{g=1}^G \sim \pi_\theta(\cdot|x)$ 
337 6:       Compute accuracy rewards:  $r_{\text{acc},g} = \text{accuracy\_score}(a_g, y^*)$ 
338 7:       Compute format rewards:  $r_{\text{fmt},g} = \text{format\_score}(\tau_g, a_g)$ 
339 8:       Phase 2: Relative High-Accuracy Trajectory Selection
340 9:       Calculate relative accuracy baseline:  $\hat{r}_{\text{acc}} = \max(\mathbb{E}_g[r_{\text{acc},g}], \gamma_{\text{acc}})$ 
341 10:      Select trajectories where  $r_{\text{acc},g} \geq \hat{r}_{\text{acc}}$ 
342 11:      Phase 3: Relative Consistency Evaluation
343 12:      for selected trajectories  $(\tau_g, a_g)$  do
344 13:        Compute reference likelihood:  $p_g = \frac{1}{|a_g|} \sum_{i=1}^{|a_g|} \pi_\phi(a_{g,i} | x, \tau_g, a_{g,<i})$ 
345 14:        Clip likelihood:  $\tilde{p}_g = \min(p_g, \gamma_p)$ 
346 15:      end for
347 16:      Calculate relative consistency baseline:  $\hat{\mu}_p = \mathbb{E}_g[\tilde{p}_g] - \epsilon_p$ 
348 17:      Select consistent trajectories where  $\tilde{p}_g \geq \hat{\mu}_p$ 
349 18:      Phase 4: Enhanced Reward Calculation
350 19:      for each trajectory  $g$  do
351 20:        
$$R_g = \underbrace{r_{\text{acc},g} + r_{\text{fmt},g}}_{\text{base reward}} + \underbrace{\lambda_{\text{cons}} \cdot r_{\text{acc},g} \cdot \mathbb{I}[\text{consistent}]}_{\text{consistency bonus}}$$

352 21:        Normalize advantage:  $\hat{A}_g = (R_g - \mu_R) / \sigma_R$ 
353 22:      end for
354 23:    end for
355 24:    Phase 5: Model Update
356 25:    Update  $\pi_\theta$  via GRPO policy gradient (without KL penalty)
357 26:    if  $t \bmod k = 0$  then ▷ EMA update every  $k = 10$  steps
358 27:       $\phi \leftarrow \alpha \phi + (1 - \alpha) \theta$  ▷  $\alpha=0.995$  typical
359 28:    end if
360 29:  end for
361 30:  return optimized policy  $\pi_\theta$ 
362 31: end procedure

```

---

near-average samples). Trajectories that exceed this baseline receive a sparse *consistency bonus*, weighted by their accuracy, ensuring that rewards prioritize both correctness and logical coherence.

**Model Update.** To promote exploration of diverse reasoning paths, we remove the KL penalty from the GRPO’s training objective. Instead, we rely on the consistency bonus—added to the base reward to form the total reward—to guide online model updates toward higher-quality outputs. The reference model is updated via EMA every few steps, allowing it to gradually inherit improvements from the online model (e.g., better visual grounding or more complex reasoning) while maintaining stability against sampling noise. This balanced optimization process enhances multimodal understanding without sacrificing logical consistency, ultimately improving both performance and interpretability.

#### 4.1 EVALUATION ON SEED-BENCH-R1

We first evaluate our method on SEED-Bench-R1. As shown in Tab. 2, GRPO-CARE significantly outperforms GRPO across all three difficulty levels, with a particularly notable improvement of nearly 10% on the most challenging L3 evaluation in domains such as *Hobbies* and *Work*, which exhibit substantial distributional divergence from the training data.

To thoroughly assess the effectiveness of GRPO-CARE, we compare it against two families of baseline methods: **KL-oriented baselines**, which modify the application of divergence constraints, and **reward-based alternatives**, which replace KL penalties with consistency-aware rewards.

- **KL-Oriented Baselines.** 1) **KL-EMA** introduces an EMA-updated reference model for adaptive constraints, following the implementation of WARP (Ramé et al., 2024). 2) **KL-EMA-HA** selectively applies KL penalty only to high-accuracy samples, applying regularization on where alignment matters most. 3) **SepKL-EMA-HA** further decomposes KL into separate terms for reasoning and answer tokens to alleviate disproportionately penalizing lengthy reasoning tokens while potentially overlooking answer-reasoning inconsistencies. 4) **NoKL** removes the KL penalty, demonstrating the raw optimization potential absent any regularization.
- **Reward-Based Alternatives.** 5) **DenseCons** applies continuous likelihood weighting to derive dense consistency rewards:  $r_{\text{cons}} = \lambda_{\text{cons}} \cdot r_{\text{acc}} \cdot p_{\phi}(a|x, \tau)$ . 6) **RefGen** takes a more explicit approach by having the reference model regenerate answers from sampled reasoning paths, using the regenerated answer’s accuracy as the consistency signal:  $r_{\text{cons}} = \lambda_{\text{cons}} \cdot \text{accuracy}(a' \sim \pi_{\phi}(\cdot|x, \tau), y^*)$ .

As shown in Tab. 3, we report both benchmark performance and the consistency rate between generated reasoning and final answers, where consistency is evaluated by GPT-4.1 to assess whether the reasoning sufficiently supports the answer. The details for consistency evaluation is described in Appendix B.3. Our analysis shows that while the EMA-updated reference model improves both accuracy and consistency, restricting KL penalties to high-accuracy samples (*KL-EMA-HA*) boosts in-domain (L1) results but slightly reduces OOD (L2/L3) generalization. Decomposing KL penalties (*SepKL-EMA-HA*) mitigates reasoning-answer inconsistency, yielding minor gains on L2 but limited impact on L3. Notably, none of the KL-based variants outperform *NoKL*, indicating that standard KL regularization may hinder the performance ceiling in this context.

Among reward-based methods, *DenseCons* surpasses *NoKL* on L1 and L2 with improved consistency, but slightly underperforms on L3, likely due to over-reliance on reference model calibration. *RefGen* greatly increases consistency but introduces instability from sampling-based answer regeneration, ultimately reducing overall performance. Our proposed **GRPO-CARE** uses sparse consistency rewards to achieve robust improvements across all levels. Its two-stage filtering—leveraging adaptive EMA-updated reference likelihoods to provide relative, sparse feedback for high-accuracy samples—effectively enhances logical consistency and answer accuracy. This demonstrates that group-relative sparse rewards deliver more reliable learning signals, avoiding overfitting to imperfect likelihoods (as in *DenseCons*) or sampling noise (as in *RefGen*).

**Effect of the EMA Reference Model.** To investigate the role of the EMA reference model in consistency assessment for reward shaping, we conduct additional analysis in Tab. 4. Using an EMA-updated reference yields significantly better performance than a frozen reference, suggesting that EMA effectively adapts the reference model to the evolving reasoning process of the online policy. Importantly, the EMA-updated reference does not simply converge to the online policy: performance with the EMA reference surpasses that of using the online policy itself, confirming that it remains a distinct and stable reference.

**Hyperparameter Sensitivity.** We use the same hyperparameters across different benchmarks: EMA update every 10 steps, EMA decay 0.995, likelihood cap 0.95, and consistency margin 0.01. We further validate the robustness of GRPO-CARE by ab-

Table 3: Ablation results of KL-oriented and reward-based baselines on SEED-Bench-R1.

| Models       | L1          | L2          | L3          | Consistency |
|--------------|-------------|-------------|-------------|-------------|
| GRPO         | 52.3        | 53.2        | 46.7        | 57.9        |
| KL-EMA       | 54.7        | 54.1        | 49.4        | 60.0        |
| KL-EMA-HA    | 55.1        | 53.8        | 49.2        | 61.7        |
| SepKL-EMA-HA | 54.8        | 54.9        | 47.5        | 76.8        |
| NoKL         | 55.4        | 54.4        | 51.3        | 70.0        |
| DenseCons    | 56.6        | 55.5        | 50.6        | 80.3        |
| RefGen       | 55.2        | 54.2        | 49.4        | <b>86.4</b> |
| CARE (ours)  | <b>57.0</b> | <b>57.0</b> | <b>53.4</b> | 82.4        |

Table 4: Impact of different reference models on GRPO-CARE.

| Reference     | L1          | L2          | L3          |
|---------------|-------------|-------------|-------------|
| Frozen        | 51.6        | 50.4        | 51.1        |
| EMA updated   | <b>57.0</b> | <b>57.0</b> | <b>53.4</b> |
| Online policy | 52.7        | 53.5        | 51.2        |

Table 5: Ablation results on key hyperparameters.

| Hyperparameter     | Values                   | Performance                      |
|--------------------|--------------------------|----------------------------------|
| Consistency margin | 0.0 / 0.01 / 0.05 / 0.1  | 56.2 / <b>57.0</b> / 56.6 / 50.5 |
| Likelihood cap     | 1.0 / 0.99 / 0.95 / 0.90 | 54.0 / 56.1 / <b>57.0</b> / 56.0 |
| EMA frequency      | 1 / 10 / 50 / $\infty$   | 54.1 / <b>57.0</b> / 56.8 / 51.6 |
| EMA decay          | 0 / 0.99 / 0.995 / 0.999 | 52.9 / 56.1 / <b>57.0</b> / 52.5 |

Table 6: Performance of different models on general video understanding benchmarks

| Models                                | VSI-Bench   | VideoMMU    | MMVU        | MVBench     | TempCompass | VideoMME    |
|---------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| GPT-4o (Hurst et al., 2024)           | 34.0        | 61.2        | 75.4        | -           | -           | 71.9        |
| LLaMA-VID (Li et al., 2024c)          | -           | -           | -           | 41.9        | 45.6        | -           |
| VideoLLaMA2 (Cheng et al., 2024)      | -           | -           | 44.8        | 54.6        | -           | 47.9        |
| LongVA-7B (Zhang et al., 2024b)       | 29.2        | 23.9        | -           | -           | 56.9        | 52.6        |
| VILA-1.5-8B (Lin et al., 2024)        | 28.9        | 20.8        | -           | -           | 58.8        | -           |
| Video-UTR-7B (Yu et al., 2025a)       | -           | -           | -           | 61.1        | 62.5        | 56.0        |
| LLaVA-OneVision-7B (Li et al., 2024a) | 32.4        | 33.8        | 49.2        | 56.7        | -           | 58.2        |
| Kangeroo-8B (Liu et al., 2024a)       | -           | -           | 61.1        | 62.5        | 69.9        | 55.4        |
| Video-R1-7B (Feng et al., 2025)       | <b>35.8</b> | <b>52.3</b> | 63.8        | 63.9        | 73.2        | 59.3        |
| Qwen2.5-VL-7B                         | 30.1        | 48.1        | 60.0        | 59.0        | 72.6        | 56.6        |
| CARE-7B (SB-R1)                       | 34.3        | 51.6        | <b>66.2</b> | 63.2        | <b>74.3</b> | 58.1        |
| CARE-7B (Video-R1)                    | <b>35.8</b> | 50.4        | 65.8        | <b>65.1</b> | 73.5        | <b>59.6</b> |

lating key hyperparameters. Table 5 reports L1 performance on SEED-Bench-R1 under different settings, confirming that our default choices are stable and effective.

## 4.2 GENERALIZATION AND TRANSFERABILITY BEYOND SEED-BENCH-R1

To comprehensively evaluate our model’s capabilities, we conduct extensive experiments on both general video understanding and language-only reasoning benchmarks beyond SEED-Bench-R1.

For video understanding, we benchmark our model on six challenging datasets spanning diverse aspects: spatial reasoning (VSI-Bench (Yang et al., 2024)), knowledge-intensive QA (VideoMMU (Hu et al., 2025) and MMVU (Zhao et al., 2025b)), and general video understanding (MVBench (Li et al., 2024b), TempCompass (Liu et al., 2024b), and VideoMME (Fu et al., 2024)). As shown in Tab. 6, our CARE-7B (SB-R1) achieves significant performance improvements over the base model across all benchmarks after training on SEED-Bench-R1. These consistent gains validate the quality of our benchmark’s training data, the robustness of our methodology, and the comprehensiveness of our evaluation protocol. Furthermore, we conduct additional experiments following Video-R1 (Feng et al., 2025), training our model using GRPO-CARE with 16-frame video inputs on general-domain data (Video-R1-260k) for 1k RL steps and testing with 32-frame inputs. The comparative results from other baselines are taken from the Video-R1 paper. Notably, even when trained solely with RL, our model achieves competitive or superior performance compared to Video-R1-7B on most benchmarks. This is particularly remarkable given that Video-R1-7B benefits from explicit temporal order grounding constraints via GRPO rewards and supplementary supervised fine-tuning with additional data. Our model’s ability to match or outperform this strong baseline with a more streamlined training pipeline underscores the efficiency of our method.

Beyond video understanding, we further evaluate whether our approach generalizes to purely language-based reasoning tasks. Specifically, we test Qwen2.5-VL-7B on GSM8k (Cobbe et al., 2021) and GPQA (Rein et al., 2024) benchmarks, comparing the base model with variants trained using GRPO and GRPO-CARE on SEED-Bench-R1. As reported in Tab. 7, GRPO-CARE yields consistent improvements over both the base model and vanilla GRPO. Interestingly, while GRPO alone slightly degrades performance on GPQA, incorporating CARE effectively reverses this trend and produces consistent gains. These results suggest that our method not only strengthens multimodal reasoning but also enhances general reasoning capabilities in purely textual domains.

Table 7: Performance on language-only reasoning benchmarks.

| Method        | GSM8k        | GPQA         |
|---------------|--------------|--------------|
| Qwen2.5-VL-7B | 72.10        | 28.79        |
| GRPO          | 74.07        | 24.24        |
| GRPO-CARE     | <b>81.58</b> | <b>30.81</b> |

## 5 CONCLUSION

In this paper, we introduced SEED-Bench-R1, a structured benchmark for evaluating post-training methods for MLLM, and proposed GRPO-CARE, a novel consistency-aware RL framework. Our analysis shows that while outcome-supervised GRPO improves accuracy, it often sacrifices reasoning coherence. GRPO-CARE addresses this by rewarding both correctness and consistency using likelihood calibration, leading to stronger generalization, higher interpretability, and effective transfer across tasks. We envision SEED-Bench-R1 and GRPO-CARE as useful tools for advancing robust post-training methods, driving the development of more powerful MLLMs.

## REFERENCES

- Learning to reason with LLMs, 2024. URL <https://openai.com/index/learning-to-reason-with-llms/>. 2, 3
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 16
- Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Alphamath almost zero: Process supervision without process. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 3
- Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*, 2023. 2, 3
- Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 9
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 9
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 16
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL <https://doi.org/10.1007/s11263-021-01531-2>. 3, 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, 2025. URL <https://arxiv.org/abs/2503.17352>. 16
- Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024. 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, 2025. URL <https://arxiv.org/abs/2503.21776>. 2, 3, 9, 15, 16
- Chaoyou Fu, Yuhao Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhan 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. 9
- Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Dayiheng Liu, Chang Zhou, Wen Xiao, et al. Llm critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. *arXiv preprint arXiv:2406.14024*, 2024. 3
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022. 3, 4

- 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. 2, 3, 4
- Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 9
- 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. 2, 3
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 9
- Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2881–2889, 2020. 3
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. 9
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 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*, pp. 22195–22206, 2024b. 3, 9
- 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*, pp. 323–340. Springer, 2024c. 9
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step, 2023. URL <https://arxiv.org/abs/2305.20050>. 3
- Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024. 9
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 16
- Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024a. 9
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 8731–8772, 2024b. 3, 9
- Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 24–32, 2022. 3
- 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*, 2025a. 3

- 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*, 2025b. 2, 3, 16
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 16
- Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024. 3
- Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 16
- 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. 2
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2, 3
- Lu Qiu, Yuying Ge, Yi Chen, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024. 2, 3
- Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedo, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. Warp: On the benefits of weight averaged rewarded policies, 2024. URL <https://arxiv.org/abs/2406.16768>. 3, 8
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>. 9
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL <https://arxiv.org/abs/1707.06347>. 3
- 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, 4
- Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1358–1375, 2024. 3
- 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, 3
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. 3
- Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Y Wu, and Zhifang Sui. Math-shepherd: A label-free step-by-step verifier for llms in mathematical reasoning. *arXiv preprint arXiv:2312.08935*, 2023. 3

- 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. 16
- Xiaodong Wang and Peixi Peng. Open-r1-video. <https://github.com/Wang-Xiaodong1899/Open-R1-Video>, 2025. 3
- Tong Wei, Yijun Yang, Junliang Xing, Yuanchun Shi, Zongqing Lu, and Deheng Ye. Gtr: Guided thought reinforcement prevents thought collapse in rl-based vlm agent training, 2025. URL <https://arxiv.org/abs/2503.08525>. 3
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 3
- Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27723–27730, 2025. 3
- Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024. 9
- Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025. 2
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 16
- En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025a. 9
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025b. URL <https://arxiv.org/abs/2503.14476>. 3
- Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 16
- Jingyi Zhang, Jiaying 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. 2, 3
- Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024a. 3
- Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b. 9
- Jiaying Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcing learning. *arXiv preprint arXiv:2503.05379*, 2025a. 3
- 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*, 2025b. 9

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingdong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL <https://arxiv.org/abs/2504.10479>. 16

## A DETAILS OF SEED-BENCH-R1

The questions from SEED-Bench-R1 are presented as multiple-choice problems and organized into a three-level hierarchy: Level 1 (in-distribution), Level 2 (cross-environment), and Level 3 (cross-task-environment). Figure 4 shows example questions from each level, where the model is required to reason about the next appropriate action using world knowledge, based on the specified task goal and the visual inputs showing task progress and current observation. Specifically, Levels 1 and 2 focus on daily-life household tasks similar to those in the training data. Level 1 questions are set in environments seen during training, while Level 2 questions are set in new, previously unseen environments. Level 3 is the most challenging, covering a broader range of task domains—including Work, Hobbies, and Recreation, as well as Daily Life—and takes place in a wider variety of unseen indoor and outdoor environments. The complete annotation files are included in the attachment, and we will make the corresponding videos publicly available after the release of this paper.

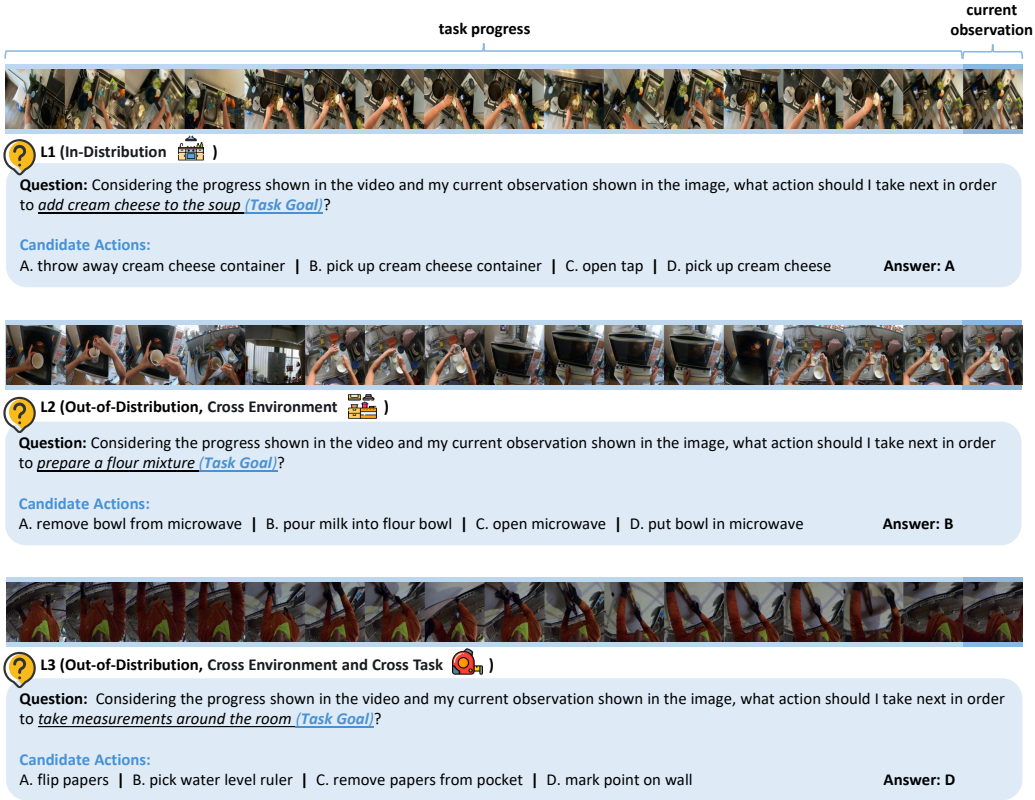


Figure 4: Example questions from the three-level evaluation hierarchy in SEED-Bench-R1’s validation set, including in-distribution, cross-environment, and cross-environment-task scenarios.

## B EXPERIMENTAL DETAILS

### B.1 COMPUTE RESOURCES

For both SFT and GRPO, we utilize four 80GB GPUs with a batch size of 4. The generation group size for GRPO is set to 8 per sample. To improve training efficiency, the number of video frames is limited to 16, with each frame resized to a resolution of  $128 \times 28 \times 28$ . In the experiments on SEED-Bench-R1, we train the model using 6k out of 50k samples from SEED-Bench-R1’s training data for the pilot study. For experiments involving Video-R1-260k, we follow the protocol in Video-R1 (Feng et al., 2025) to train the model for 1,000 steps using a subset of the data. The training time is about 8 hours for GRPO and 50 minutes for SFT.

## B.2 ADDITIONAL EVALUATION ON DIFFERENT BASE MODELS

To further validate the generality of our method, we conduct experiments with GRPO-CARE on multiple Qwen variants of different scales. Recent RL-based post-training studies for MLLMs (Liu et al., 2025b; Feng et al., 2025; Deng et al., 2025) commonly adopt Qwen series models due to their strong open-source ecosystem and competitive performance on multimodal benchmarks. Following this trend, we select Qwen2.5-VL-3B and Qwen2-VL-7B as alternative base models in our evaluation.

As shown in Tab. 8, GRPO-CARE consistently surpasses both SFT and vanilla GRPO across all tested variants. The performance improvements are not only observed on larger models like Qwen2-VL-7B, but also on smaller-scale ones such as Qwen2.5-VL-3B, indicating that our approach is both robust and scalable.

For context, we also report the performance of several representative MLLMs on SEED-Bench-R1 in Tab. 8. These results serve as reference baselines to illustrate the overall performance range of current models. Notably, GRPO-CARE consistently improves Qwen variants beyond their supervised and vanilla GRPO counterparts, demonstrating potential for extension to broader multimodal architectures.

## B.3 CONSISTENCY EVALUATION

We use GPT-4.1 to assess the consistency between the model’s reasoning process and its final answer. The detailed prompt is shown in Figure 5, where GPT-4.1 is instructed to analyze the logical flow of the model’s response and assign a score reflecting whether the reasoning supports the final answer.

```

Question:{}
Procedure:{}
Answer:{}

The above is a Question, along with a model's Solution Procedure and the Answer.

Please fulfill the following requirements:
1. Consistency Analysis. Check if the solution process supports the answer and if the answer can be logically derived from the process. If consistent, the score is 1, with reasons provided.
2. Logic Flow Analysis. For example, derive A to get answer X; refine step A to verify the correctness of answer X.
3. Other Analysis. Identify any other issues in the solution steps and name them to alert the solver.

Reply in JSON format:
```json
{"consistency_analysis":{"score": 0 or 1, "reasons":""}, "logic_flow":"analysis of procedure", "others":""}
```

There is no need to consider whether the answer is correct; just analyze the above indicators.

```

Figure 5: The detailed prompt for evaluating consistency between the reasoning process and the final answer using GPT-4.1.

To validate GPT-4.1’s judgments, we randomly sampled 50 consistent and 50 inconsistent cases flagged by GPT-4.1 and compared them with human annotations, achieving a Pearson correlation of 0.82. This demonstrates that GPT-4.1 provides reliable and nuanced evaluations of consistency. Below we present example GPT-4.1 evaluations to illustrate the evaluation process.

Table 8: Performance of various MLLMs on SEED-Bench-R1. GRPO-CARE consistently improves performance across different base models and model sizes.

| Model                               | L1   | L2   | L3   |
|-------------------------------------|------|------|------|
| BLIP-2 (Li et al., 2023)            | 26.4 | 27.3 | 26.2 |
| InstructBLIP (Dai et al., 2023)     | 27.0 | 25.9 | 26.3 |
| Valley (Luo et al., 2023)           | 26.8 | 26.8 | 27.0 |
| Yi-VL (Young et al., 2024)          | 30.0 | 29.9 | 23.5 |
| LLaVA1.5 (Liu et al., 2023)         | 30.7 | 31.0 | 25.4 |
| DeepSeek-VL (Lu et al., 2024)       | 31.5 | 33.2 | 28.5 |
| VideoLLaMA3-7B (Zhang et al., 2023) | 33.3 | 33.2 | 27.7 |
| InternVL3-8B (Zhu et al., 2025)     | 38.0 | 37.8 | 31.8 |
| Qwen2-VL-7B (Wang et al., 2024)     | 34.7 | 34.0 | 31.6 |
| Qwen2-VL-7B + SFT                   | 43.8 | 44.1 | 38.2 |
| Qwen2-VL-7B + GRPO                  | 46.0 | 50.2 | 44.9 |
| Qwen2-VL-7B + GRPO-CARE             | 57.2 | 56.2 | 53.8 |
| Qwen2.5-VL-3B (Bai et al., 2025)    | 31.3 | 32.7 | 28.2 |
| Qwen2.5-VL-3B + SFT                 | 35.9 | 39.1 | 33.7 |
| Qwen2.5-VL-3B + GRPO                | 39.6 | 41.0 | 35.4 |
| Qwen2.5-VL-3B + GRPO-CARE           | 47.1 | 48.8 | 43.5 |

**Example 1.**

```

{
  "task_goal": "prepare the teapot with tea leaves",
  "options": [
    "A. put down teabag holder",
    "B. scoop tea leaves",
    "C. take bag of tea",
    "D. open bag of tea"
  ],
  "model_response": "<think>Let me think. The person has already opened
    the bag of tea and is now reaching for the teapot. This suggests
    that they are in the process of preparing the teapot with tea
    leaves. The next logical step would be to scoop the tea leaves into
    the teapot. Therefore, the most appropriate action would be to
    scoop the tea leaves.</think><answer>B</answer>",
  "gpt_evaluation": {
    "consistency_analysis": {
      "score": 1,
      "reasons": "The procedure describes reaching for the teapot after
        opening the bag of tea, and logically concludes that the next
        action should be to scoop tea leaves. This directly supports
        the chosen answer (B: scoop tea leaves).",
    },
    "logic_flow": "The procedure starts from the current observation (bag
      of tea opened, reaching for teapot), and infers that scooping
      tea leaves is the next step in preparing the teapot, which
      matches the answer given. Each reasoning step directly leads to
      the selected answer.",
    "others": "No other major issues identified. The reasoning is clear
      and follows from the initial observations and procedural logic."
  }
}

```

**Example 2.**

```

{
  "task_goal": "wash knife, cutting board, and peeler",
  "options": [
    "A. put cutting board onto drying rack",
    "B. put knife onto drying rack",
    "C. pick up cutting board",
    "D. adjust tap"
  ],
  "model_response": "<think>The person has just adjusted the tap, which
    suggests they may have finished washing the cutting board. The
    cutting board is currently resting on the drying rack, which
    implies it may be ready to be washed.</think><answer>A</answer>",
  "gpt_evaluation": {
    "consistency_analysis": {
      "score": 0,
      "reasons": "The procedure states that the cutting board is already
        resting on the drying rack, which implies that action A (put
        cutting board onto drying rack) is redundant or already
        completed. Therefore, the process does not support the answer,
        and it cannot be logically derived from the provided reasoning
        .",
    },
    "logic_flow": "The reasoning chain concludes that the cutting board
      is already on the drying rack, so logically the next step should
      not be to put it there again. Instead, attention should shift to
      another item (knife or peeler) or another action. This breaks the
      logic flow from observation to answer.",
    "others": "There is confusion in the think step: it simultaneously
      claims that the cutting board is on the drying rack and that it

```

```

    may be ready to be washed, which are contradictory states.
    Clearer distinction between current and next state is needed."
  }
}

```

## C LIMITATIONS

While our work introduces SEED-Bench-R1 and the GRPO-CARE framework to advance post-training for MLLMs, several limitations remain. (1) Although SEED-Bench-R1 provides a rigorous benchmark for video understanding with hierarchical evaluation, it does not yet encompass all possible multimodal domains or modalities. Expanding its coverage in future iterations will be important for further validating the generality of models. (2) The consistency-aware reward in GRPO-CARE relies on model-internal likelihoods and group calibration, which, despite their effectiveness, may not fully capture subtle reasoning errors or always align with human judgment. (3) Similar to other RL-based frameworks, our method incurs additional computational costs due to the need for maintaining reference models and performing group-based calibration. While this overhead was manageable in our experiments, scaling to larger models or more complex tasks may necessitate further optimization. We believe these aspects open valuable opportunities for further research built upon our contributions.

## D BROADER IMPACTS

On the positive side, the proposed consistency-aware reinforcement learning framework, GRPO-CARE, enhances transparency in AI systems by encouraging logically coherent reasoning chains. This is particularly important in domains such as education, healthcare, and assistive robotics, where understanding the reasoning process is essential. Meanwhile, by mitigating shortcut learning, our approach improves generalization to real-world scenarios, paving the way for safer and more reliable AI applications. Moreover, the release of SEED-Bench-R1 alongside our training framework provides valuable tools to support ongoing research in multimodal understanding.

Nonetheless, potential negative societal impacts warrant careful consideration. Although our method enhances reasoning consistency, it does not directly address biases in training data, leaving the risk of perpetuating or amplifying existing societal biases. Additionally, greater reliance on automated reasoning—even when more interpretable—may lead to overconfidence in AI outputs and diminished human oversight. Finally, the substantial computational resources required to train and evaluate large MLLMs raise environmental concerns due to increased energy consumption.

## E USE OF LARGE LANGUAGE MODELS

We used large language models (LLMs) solely for language editing and polishing purposes. The LLMs were not involved in the design of research ideas, development of methods, data analysis, or interpretation of results. All conceptual contributions, technical content, and scientific claims in this paper are entirely the work of the authors.