GRPO-CARE: Consistency-Aware Reinforcement Learning for Multimodal Reasoning

Yi Chen[1], **Yuying Ge**[2*], **Rui Wang**[3], **Yixiao Ge**[2*], **Junhao Cheng**[2], **Ying Shan**[2], **Xihui Liu**[1*]
[1]The University of Hong Kong, [2]ARC Lab, Tencent PCG, [3]The Chinese University of Hong Kong
https://github.com/TencentARC/GRPO-CARE

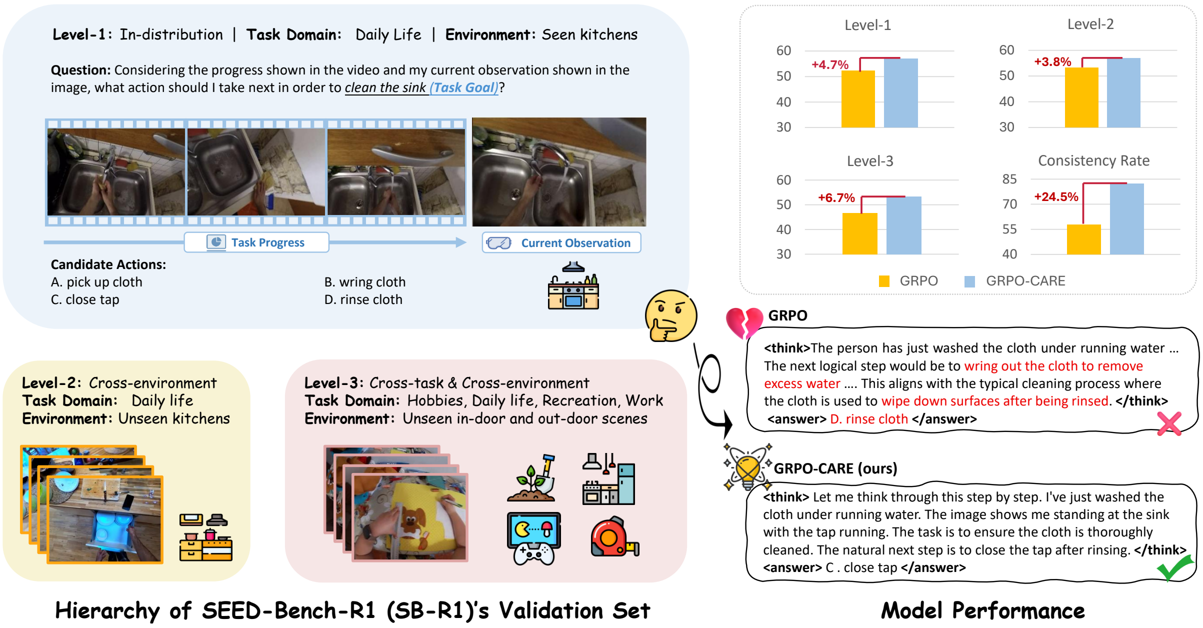

Figure 1: (a) **SEED-Bench-R1** (SB-R1) provides a systematic, three-level evaluation of post-training methods for MLLMs, encompassing tasks that require both perception and reasoning to tackle complex real-world scenarios. (b) Our analysis identifies a key limitation of standard outcome-supervised GRPO: 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 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 present **SEED-Bench-R1**, a structured testbed featuring real-world video tasks and hierarchical evaluation across in-distribution, cross-environment, and cross-environment-task scenarios. Our analysis reveals that standard outcome-supervised GRPO often yields "logical incoherence"—achieving correct answers through flawed reasoning—due to its exclusive focus on final-answer rewards and rigid KL penalties. To address this, we propose **GRPO-CARE**, a consistency-aware RL framework that eliminates KL penalties while introducing a two-tiered reward system: a base reward for accuracy and an adaptive bonus for consistency. This bonus, derived from a slowly evolving reference model through group-relative likelihood calibration, rewards reasoning paths that logically support the final answer without requiring expensive process supervision. 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, validating its robustness and generality.

*Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 4305–4320
July 2-7, 2026 ©2026 Association for Computational Linguistics

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 (Liu et al., 2025b), others on logical reasoning (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 levels*. 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 (out-of-distribution) 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 **C**onsistency-**A**ware **R**eward **E**nhancement 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.

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

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.

2 Related Work

RL for LLMs/MLLMs. RLHF aligns LLMs with human preferences (Ouyang et al., 2022; Schulman et al., 2017), while long CoT generation significantly boosts complex reasoning (Guo et al., 2025; Team et al., 2025; ope, 2024). Outcome-based RL, such as GRPO (Shao et al., 2024) and its variants (Yu et al., 2025b; Liu et al., 2025a), optimizes CoT but often suffers from inconsistent reasoning. Previous remedies include training process reward models with step-wise annotations (Lightman et al., 2023; Uesato et al., 2022; Chen et al.; Luo et al., 2024; Wang et al., 2023), employing LLM judges (Gao et al., 2024; Xia et al., 2025; Zhang et al., 2024a), or using EMA-updated reference models (Ramé et al., 2024). In MLLMs, outcome-only RL can trigger "thought collapse," which is typically mitigated by stronger correctors (Wei et al., 2025) or reward matching (Zhang et al., 2025). Crucially, unlike existing approaches like WARP (Ramé et al., 2024) that apply an EMA model merely as a passive KL-penalty constraint, our *GRPO-CARE* utilizes a slowly updated reference model to actively shape the reward signal. By performing group-relative likelihood calibration, it generates a consistency bonus that directly incentivizes logical reasoning, eliminating the need for additional annotations or external evaluation models.

Benchmarks for MLLM Post-training. Current RL-based post-training primarily focuses on image-level perception and reasoning (Huang et al., 2025; Liu et al., 2025b; Zhang et al., 2025; Sun et al., 2024), leaving video understanding relatively underexplored. Existing video RL efforts are limited by narrow task scopes (Wang and Peng, 2025; Zhao et al., 2025a; Liu et al., 2022; Jiang et al., 2020) or insufficient data for scalable training (Wu et al., 2024). Furthermore, general video benchmarks (Li et al., 2024b; Liu et al., 2024b; Fang et al., 2024; Feng et al., 2025) often lack systematic assessments of model generalization. We present *SEED-Bench-R1*, a comprehensive benchmark providing: (1) large-scale post-training data, (2) structured validation for multi-level generalization, and (3) a balanced evaluation of multimodal perception and reasoning in real-world scenarios.

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 of daily activities (Damen et al., 2022; Grauman et al., 2022). 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, to discern the correct action plan. Moreover, the derivation of golden answers is traceable and easy to verify.

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

Models	L1 (In-Distribution)	L2 (Cross-Env)	L3 (Cross-Task, Cross-Env)				
	Daily Life	Daily Life	Daily Life	Hobbies	Recreation	Work	Overall
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)	**57.0**	**57.0**	**57.6**	**51.2**	**57.4**	**48.5**	**53.4**

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

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 downsampled to 16 frames with a maximum of 128 patches, each having a spatial resolution of 28×28 pixels, 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 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 τ_g and a_g denote reasoning and answer. Unlike SFT, GRPO does not rely on predefined responses. The policy is optimized by maximizing:

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

where ε and β are hyperparameters, and $\mathbb{D}_{KL}$ is the KL divergence between policy π_θ and reference π_{ref}. The per-token advantage is set to the normalized reward $\hat{A}_{g,i} = \widetilde{r}_g = \frac{r_g - \text{mean}(\{r_1,\ldots,r_G\})}{\text{std}(\{r_1,\ldots,r_G\})}$, 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 post-trained with various methods on SEED-Bench-R1. Notably, compared to SFT, RL with GRPO significantly boosts MLLM performance on both in-distribution (L1) and OOD (L2, L3) questions, despite relying solely 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. 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

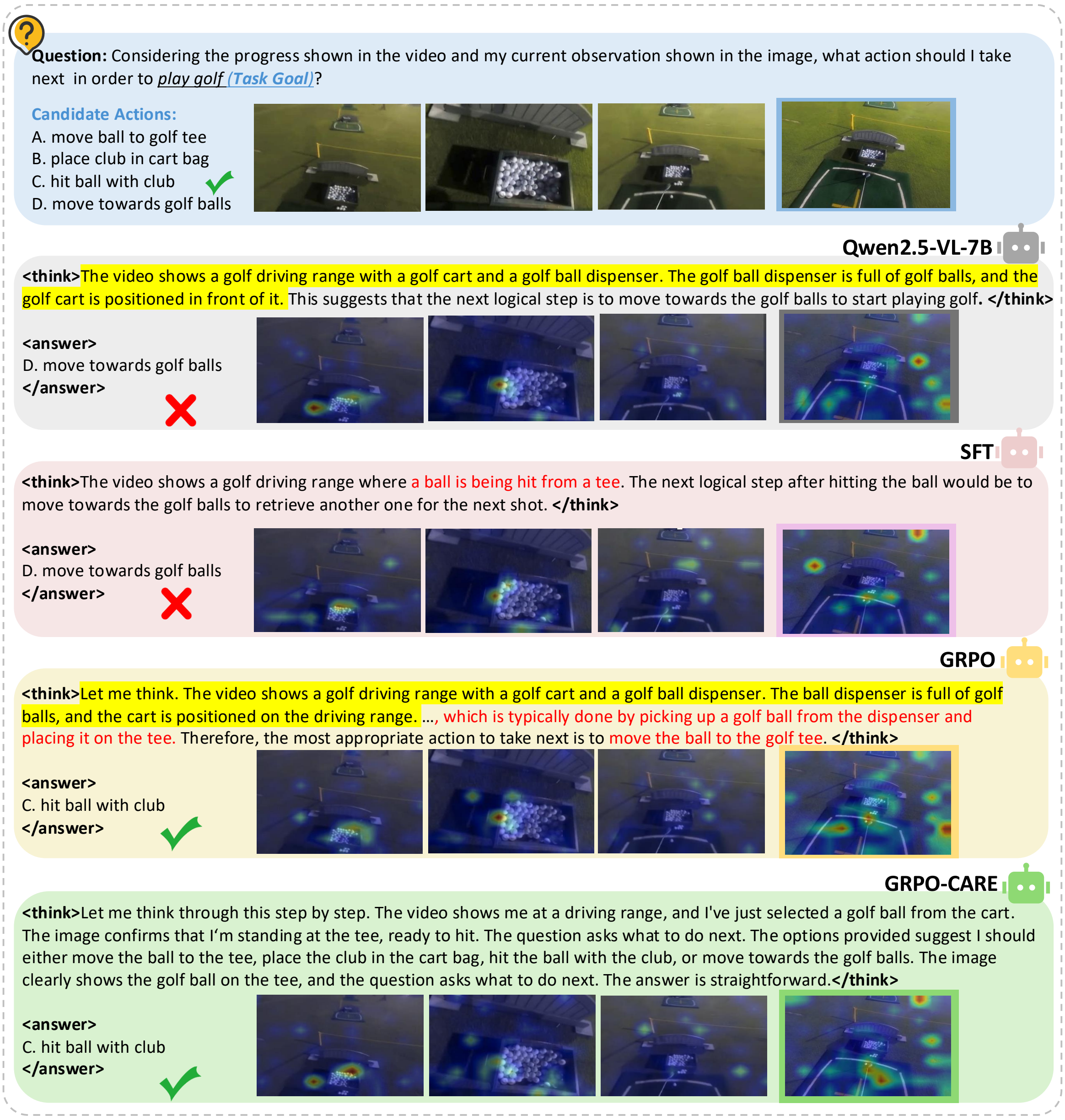

Figure 2: Case study of an L3 question from SEED-Bench-R1, showing a video of task progress, a final observation image, and attention maps (output-to-visual tokens). The SFT model tends to memorize reasoning patterns and exhibits perceptual hallucinations. The GRPO model attends more comprehensively 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.

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 reasoning is crucial for integrating world knowledge with perception. For example, in Fig. 1, the GRPO model identifies "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 interme-

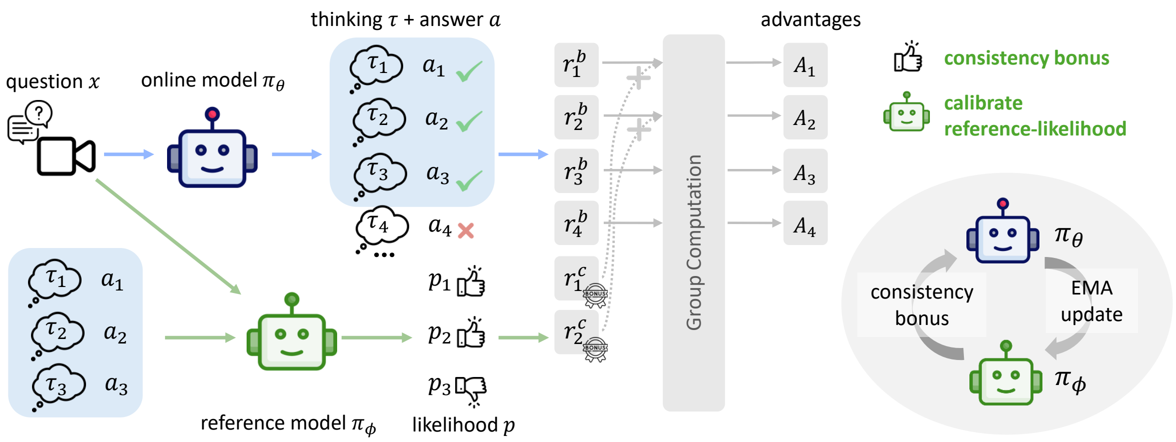

Figure 3: GRPO-CARE uses a two-tier reward system: a base reward for answer correctness (r_*^b) and an adaptive consistency bonus (r_*^c). The consistency bonus is given to high-accuracy samples whose reasoning-to-answer likelihood—estimated by a slowly updated (EMA) reference model—is higher than that 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.

diate 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** (**C**onsistency-**A**ware **R**eward **E**nhancement), 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 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). Theoretically, likelihood calibration $p_\phi(a|x, \tau)$ evaluates the logical support a reasoning path τ provides for the answer a, given the multimodal question x. A frozen reference model often lacks the domain-specific sensitivity required to evaluate the increasingly sophisticated reasoning chains produced during training. Conversely, the online policy is prone to self-confirmation bias. Our EMA model acts as a stable, lagged "teacher" controlled by a decay parameter α (e.g., $\alpha = 0.995$): it inherits the evolving reasoning capabilities of the online policy while remaining robust against training noise. This provides a more objective and consistent reward signal and ensures better gradient quality without relying on a stronger external teacher, unlike methods that depend on fixed 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 maintain calibrated confidence and prevent over-optimization toward narrow reasoning paths that might sacrifice overall accuracy.

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 near-average samples). Trajectories that exceed this baseline receive a sparse *consistency bonus*,

Algorithm 1 Consistency-Aware Reward Enhanced GRPO

Require:
$\boldsymbol{\pi_\theta}$: Online policy; $\boldsymbol{\pi_\phi}$: Reference model ($\phi \leftarrow \theta$ initially); $\boldsymbol{\mathcal{D}}$: Multimodal training dataset $\{(x, y^*)\}$
$\boldsymbol{\lambda}_{\text{cons}}$: Consistency reward coefficient (e.g., 0.5); $\boldsymbol{\gamma}_{\text{acc}}$: Minimum accuracy threshold (e.g. 0.1)
$\boldsymbol{\gamma_p}$: Maximum likelihood threshold (e.g. 0.95); $\boldsymbol{\epsilon_p}$: Consistency margin (e.g. 0.01)

procedure TRAINING($\pi_\theta, \mathcal{D}, T$)
  for $t \leftarrow 1$ to T **do**
    for each multimodal input x in batch $\mathcal{D}$ **do**
      Phase 1 (Trajectory Generation & Reward Computation):
        Generate G reasoning traces + answers: $\{\tau_g, a_g\}_{g=1}^{G} \sim \pi_\theta(\cdot|x)$
        Compute accuracy & format rewards: $r_{\text{acc},g} = \text{acc}(a_g, y^*)$ and $r_{\text{fmt},g} = \text{fmt}(\tau_g, a_g)$
      Phase 2 (Relative High-Accuracy Trajectory Selection):
        Calculate relative accuracy baseline: $\hat{r}_{\text{acc}} = \max(\mathbb{E}_g[r_{\text{acc},g}], \gamma_{\text{acc}})$; Select trajectories $\mathcal{G}_{\text{sel}} = \{g \mid r_{\text{acc},g} \geq \hat{r}_{\text{acc}}\}$
      Phase 3 (Relative Consistency Evaluation):
        for $g \in \mathcal{G}_{\text{sel}}$ **do** Compute & Clip likelihood: $p_g = \frac{1}{|a_g|}\sum_{i=1}^{|a_g|} \pi_\phi(a_{g,i} \mid x, \tau_g, a_{g,<i})$; $\tilde{p}_g = \min(p_g, \gamma_p)$
        Calculate relative consistency baseline: $\hat{\mu}_p = \mathbb{E}_g[\tilde{p}_g] - \epsilon_p$ and identify consistent trajectories if $\tilde{p}_g \geq \hat{\mu}_p$
      Phase 4 (Enhanced Reward Calculation): **for** each trajectory g **do**
        $R_g = \underbrace{r_{\text{acc},g} + r_{\text{fmt},g}}_{\text{base reward}} + \underbrace{\lambda_{\text{cons}} \cdot r_{\text{acc},g} \cdot \mathbb{I}[\text{consistent}]}_{\text{consistency bonus}}$
        Normalize advantage: $\hat{A}_g = (R_g - \mu_R)/\sigma_R$
    end for
    Phase 5 (Model Update): Update π_θ via GRPO policy gradient (without KL penalty)
      if $t \bmod k = 0$ **then** $\phi \leftarrow \alpha\phi + (1-\alpha)\theta$ ▷ EMA update every $k = 10$ steps; α=0.995 typical
  end for
  return optimized policy π_θ
end procedure

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. The ablations on key hyperparameters are presented in Appendix B.3. We further compare GRPO-CARE against two baseline families:

KL-Oriented Baselines. *1) KL-EMA* uses an EMA-updated reference model for adaptive KL constraints, following WARP (Ramé et al., 2024). *2) KL-EMA-HA* regularizes only high-accuracy samples, applying KL penalty on where alignment matters most. *3) SepKL-EMA-HA* further decomposes KL into separate reasoning and answer terms to alleviate disproportionately penalizing lengthy reasoning tokens while potentially overlooking answer-reasoning inconsistencies. *4) NoKL* removes KL penalties to test raw optimization potential.

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* uses the reference model to regenerate answers from sampled reasoning paths, using regenerated accuracy as a consistency signal: $r_{\text{cons}} = \lambda_{\text{cons}} \cdot \text{accuracy}(a' \sim \pi_\phi(\cdot|x, \tau), y^*)$.

Models	L1	L2	L3	Consistency
GRPO	52.3	53.2	46.7	57.9
KL-EMA (WARP-based)	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	**86.4**
CARE (ours)	**57.0**	**57.0**	**53.4**	82.4

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

We report both benchmark performance and

Models	VSI-Bench	VideoMMMU	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)	**35.8**	**52.3**	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	**66.2**	63.2	**74.3**	58.1
CARE-7B (Video-R1)	**35.8**	50.4	65.8	**65.1**	73.5	**59.6**

Table 4: Performance of different models on general video understanding benchmarks

Reference Model	L1	L2	L3
Frozen	51.6	50.4	51.1
EMA updated	**57.0**	**57.0**	**53.4**
Online policy	52.7	53.5	51.2

Table 5: Impact of reference models on GRPO-CARE.

Method	GSM8k	GPQA
Qwen2.5-VL-7B	72.10	28.79
GRPO	74.07	24.24
GRPO-CARE	**81.58**	**30.81**

Table 6: Performance on language-only benchmarks.

reasoning-answer consistency rate in Tab. 3. Consistency is evaluated by GPT-4.1 (details in Appendix B.4). Our analysis shows that while the EMA-updated reference model improves both accuracy and consistency (*KL-EMA*), 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, *NoKL* outperforms all KL variants, suggesting standard KL regularization may limit the optimization ceiling.

Among reward-based methods, *DenseCons* improves L1/L2 consistency but underperforms on L3, likely due to over-reliance on reference model calibration. *RefGen* boosts consistency but introduces instability from sampling-based answer regeneration, reducing overall performance. **GRPO-CARE** achieves robust gains across all levels. Its two-stage filtering—leveraging adaptive EMA-updated likelihoods for relative, sparse feedback—effectively enhances logical consistency without overfitting to imperfect likelihoods (as in *DenseCons*) or sampling noise (as in *RefGen*).

Effect of the EMA Reference Model. Tab. 5 analyzes the role of the EMA reference model in consistency assessment for reward shaping. 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.

4.2 Generalization 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.

Video Understanding. Beyond SEED-Bench-R1, we evaluate our model on six challenging datasets spanning diverse aspects: spatial reasoning (VSI-Bench (Yang et al., 2024)), knowledge-intensive QA (VideoMMMU (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)). For MMVU, we employ multiple-choice questions to ensure evaluation stability, while for VideoMME, we adopt the subtitle-free setting to focus on visual understanding. As shown in Tab. 4, 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.

Language-only Reasoning. We further examine 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), comparing the base model with variants trained using GRPO and GRPO-CARE on SEED-Bench-R1. As reported in Tab. 6, GRPO-CARE yields improvements over both the base model and vanilla GRPO. 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.

5 Conclusion

In this paper, we introduced SEED-Bench-R1, a structured benchmark for evaluating post-training methods for MLLMs, 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.

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.

References

2024. Learning to reason with LLMs.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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*.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. 2023. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

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.

2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55.

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

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Dayiheng Liu, Chang Zhou, Wen Xiao, and 1 others. 2024. Llm critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. *arXiv preprint arXiv:2406.14024*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *ICML*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024c. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024a. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024b. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8731–8772.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. 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*, pages 24–32.

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

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

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, and 1 others. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Lu Qiu, Yuying Ge, Yi Chen, Yixiao Ge, Ying Shan, and Xihui Liu. 2024. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*.

Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedoz, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. 2024. Warp: On the benefits of weight averaged rewarded policies. *Preprint*, arXiv:2406.16768.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1358–1375.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Y Wu, and Zhifang Sui. 2023. Math-shepherd: A label-free step-by-step verifier for llms in mathematical reasoning. *arXiv preprint arXiv:2312.08935*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Tong Wei, Yijun Yang, Junliang Xing, Yuanchun Shi, Zongqing Lu, and Deheng Ye. 2025. Gtr: Guided thought reinforcement prevents thought collapse in rl-based vlm agent training. *Preprint*, arXiv:2503.08525.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, and 1 others. 2025a. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*.

Qiying 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, and 16 others. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

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

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024a. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024b. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025a. R1-omni: Explainable omni-multimodal emotion recognition with reinforcing learning. *arXiv preprint arXiv:2503.05379*.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, and 1 others. 2025b. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*.

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, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

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.

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 Evaluation with 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.

Table 7: 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

As shown in Tab. 7, 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. 7. 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 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 ablating key hyperparameters. Table 8 reports L1 performance on SEED-Bench-R1 under different settings, confirming that our default choices are stable and effective.

Table 8: Ablation results on key hyperparameters.

Hyperparameter	Values	Performance
Consistency margin	0.0 / 0.01 / 0.05 / 0.1	56.2 / **57.0** / 56.6 / 50.5
Likelihood cap	1.0 / 0.99 / 0.95 / 0.90	54.0 / 56.1 / **57.0** / 56.0
EMA frequency	1 / 10 / 50 / ∞	54.1 / **57.0** / 56.8 / 51.6
EMA decay	0 / 0.99 / 0.995 / 0.999	52.9 / 56.1 / **57.0** / 52.5

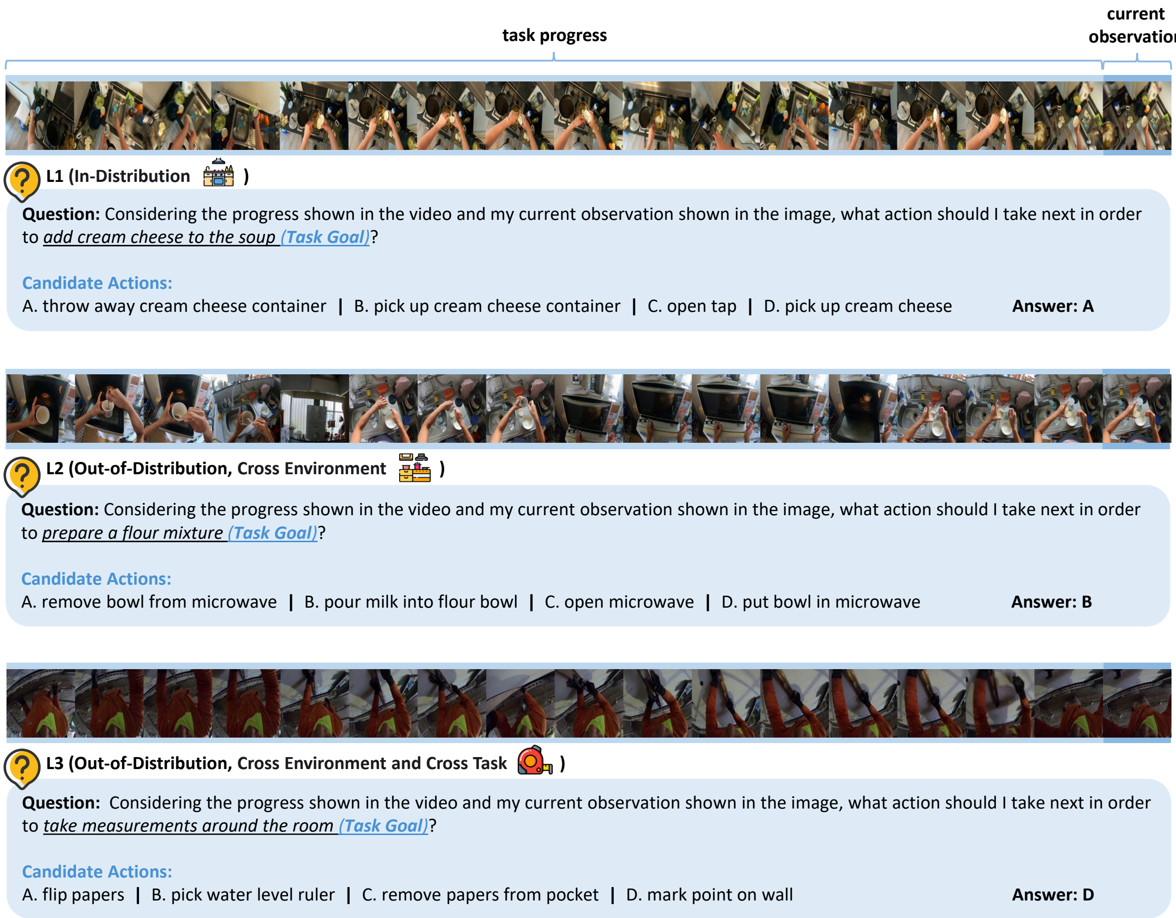

Figure 4: Example questions from the three-level evaluation hierarchy in SEED-Bench-R1's validation set, including in-distribution, cross-environment, and cross-environment-task scenarios.

```
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.

B.4 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.

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.

C 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.

D 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.

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."
  }
}
```