Robust Multimodal Reasoning via Permutation-Invariant Reinforcement Learning

Anonymous ACL submission

Abstract

We introduce Answer-Consistent REinforcement Learning (ACRE), a GRPO-compatible reward shaping method that enforces consistency between the chain-of-thought reasoning and the final answer in multimodal large language models. Outcome-only reinforcement learning improves answer accuracy but can decouple the reasoning trace from the decision: on the MMVU benchmark, standard GRPO achieves only 79.7% consistency between its generated reasoning and chosen answer, producing correct reasoning with wrong answers and vice versa. ACRE addresses this by shuffling answer options and re-prompting the model with its own reasoning trace; a consistency-verification reward grants credit only when the original and post-shuffle answers agree and are correct, penalizing reliance on option-order biases and reasoning–answer drift. Across five video reasoning and multimodal math benchmarks, ACRE yields +2.2 and +1.5 point average improvements over GRPO for video and math tasks, respectively, while using only 9.2k RL-stage training samples—28× fewer than Video-R1-260k. ACRE also matches or exceeds SFT-level measured reasoning–answer consistency on the evaluated settings and substantially improves option-shuffling robustness, achieving 74.6% shuffling consistency on MMVU versus 49.8% for GRPO.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has emerged as a powerful paradigm for enhancing the reasoning abilities of Large Language Models (LLMs). By directly optimizing for outcome correctness, approaches such as DeepSeek-R1 (Guo et al., 2025) achieve strong gains in domains where final answers are easily verified, including mathematics and coding.

This success has spurred extensions to Multimodal Large Language Models (MLLMs). Visual-RFT (Liu et al., 2025) applies reinforcement fine-tuning to visual perception tasks, Video-R1 (Feng et al., 2025) adapts Group Relative Policy Optimization (GRPO) to video understanding with temporal-aware rewards, and Vision-R1 (Huang et al., 2025) uses GRPO with hard formatting rewards on multimodal math data.

Despite these advances, a fundamental question remains under-explored: *does outcome-only RL actually improve the reasoning process, or does it merely improve the answer?* We systematically diagnose this problem in multimodal settings, revealing a pervasive *reasoning–answer inconsistency* in GRPO-trained MLLMs. We identify two failure modes: the model generates correct reasoning but selects the wrong answer (*CR–WA*), or produces flawed reasoning yet arrives at the correct answer by chance (*WR–CA*). As illustrated in Fig. 1, analyzing Video-R1-7B on MMVU reveals that 18.4% and 2.5% of samples exhibit CR–WA and WR–CA, respectively. These failures matter because GRPO assigns a trajectory-level advantage to the entire sampled response based on final-answer reward. In CR–WA cases, final-answer failure gives the whole trajectory a negative advantage, so sound reasoning tokens are suppressed together with the wrong answer. In WR–CA cases, final-answer success gives the whole trajectory a positive advantage, so flawed reasoning, option-index priors, positional heuristics, or visual shortcuts can be reinforced because they happened to yield the right letter. Together, these two modes amplify rationale–answer decoupling under outcome-only rewards, undermining the trustworthiness and interpretability of the model. Critically, this decoupling means that the chain-of-thought—intended to make model decisions interpretable—becomes unreliable as an explanation, even when the final answer is correct.

To address this, we propose ACRE (Answer-Consistent REinforcement Learning), a GRPO-compatible reward shaping scheme that explicitly

promotes consistency between the reasoning process and the final answer. For each training sample, ACRE first generates a chain-of-thought and answer as usual; it then shuffles the multiple-choice options and, conditioning on the *same* reasoning trace, prompts the model for a second answer. A maximal reward is granted only when both answers agree and are correct; otherwise, a lower reward is assigned. This consistency-verification signal penalizes reasoning–answer misalignment and discourages reliance on option-order biases, encouraging more grounded and robust multimodal reasoning.

Our contributions are as follows: **(1)** We systematically diagnose reasoning–answer inconsistency in GRPO-trained MLLMs. Through two complementary metrics—CoT Answer Consistency Rate (CACR) and Option Shuffling Consistency Rate (OSCR)—we reveal that outcome-only RL *degrades* the measured consistency between reasoning traces and final answers relative to supervised fine-tuning, even as it improves accuracy. **(2)** We propose ACRE, a GRPO-compatible reward shaping scheme that enforces shuffle-invariant agreement conditioned on correctness. By re-prompting the model with its own reasoning under permuted options, ACRE provides a self-consistency signal that requires no learned step-level verifier during training or step-level annotation, yet effectively penalizes both shortcut reliance and reasoning–answer drift. **(3)** Across five video and math reasoning benchmarks, ACRE achieves +2.2 and +1.5 average point gains over GRPO for video and math tasks respectively, while matching or exceeding SFT-level CACR on the evaluated settings and using only 9.2k RL-stage training samples—28$\times$ fewer than Video-R1-260k. We further demonstrate generalization to open-ended QA via a mixed-reward strategy.

2 Related Work

Reinforcement Learning for Multimodal Large Language Models. RL-based alignment of LLMs has progressed from preference-based methods such as DPO (Rafailov et al., 2023) and ORPO (Hong et al., 2024), which replace the RLHF (Bai et al., 2022; Ouyang et al., 2022) reward model with closed-form objectives, to verifiable or outcome-based rewards that score answers with programmatic checks. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has become a widely adopted instantiation, demonstrating strong reasoning gains in text-only LLMs and, more recently, in multimodal settings. However, pure outcome-based rewards can introduce unexpected behavior: prior work has documented mismatches between reasoning traces and final answers in LLMs (Lanham et al., 2023; Turpin et al., 2023). A complementary line of research addresses this via *process supervision*, enforcing correctness of intermediate reasoning steps. Step-level verification was shown to reduce logical errors in text-based reasoning (Lightman et al., 2024), and this paradigm has been extended to multimodal models: VisualPRM (Wang et al., 2025) uses fine-grained rewards to detect hallucinations, while Video-SALMONN-o1 (Sun et al., 2025) integrates step-level supervision into preference optimization via pDPO. Our work takes a complementary approach: rather than supervising individual reasoning steps—which requires costly step-level annotations or a trained verifier—we test whether the same full reasoning trace can recover the same answer identity after option-surface perturbation. This self-consistency signal requires no step-level annotation, requires no learned step-level verifier during training, and integrates directly into the GRPO objective as a lightweight reward shaping term.

Multimodal Reasoning Large Language Models. Chain-of-Thought (CoT) prompting (Wei et al., 2022) and supervised fine-tuning on rationale-augmented corpora (Lu et al., 2022) have become primary strategies for eliciting step-by-step reasoning in MLLMs. However, static CoT follows a rigid, unidirectional inference path that lacks mechanisms for self-correction and reflection. Recent work addresses this limitation through iterative refinement frameworks, such as Self-Refine (Madaan et al., 2023), which enables models to critique and revise their own outputs, and tool-augmented approaches like Chameleon (Lu et al., 2023b), which allow models to seek external information for intermediate verification. Our work sits at the intersection of these lines: rather than relying solely on SFT-style CoT supervision or outcome-only RL, we *shape* the RL objective with an explicit *consistency signal* that ties the final answer to the reasoning trace.

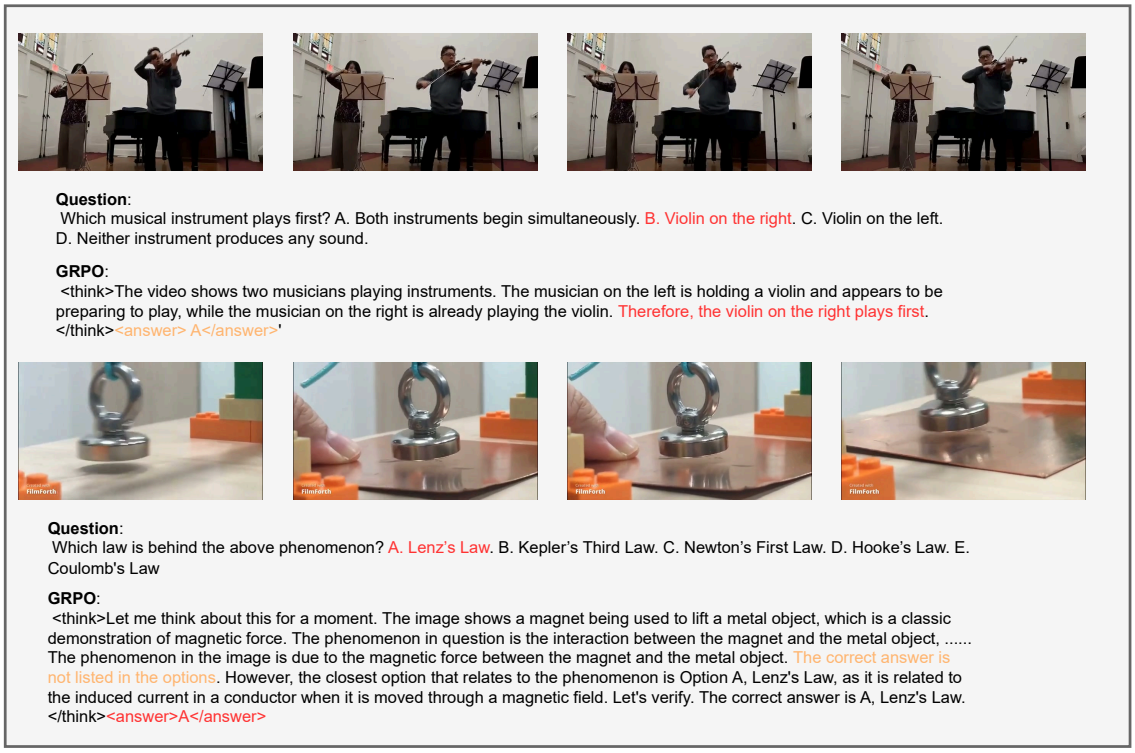


Figure 1: Reasoning–answer inconsistency of GRPO models. Red denotes correct answers or reasoning traces, and orange denotes flawed answers or reasoning traces. The upper example shows Correct Reasoning but Wrong Answer (CR–WA); the lower example shows Wrong Reasoning but Correct Answer (WR–CA).

3 Reasoning-Answer Inconsistency in Multimodal Large Language Models

We study reasoning–answer inconsistency in post-trained multimodal LLMs. Let M_r denote a reasoning MLLM that, given a multimodal input x and a text query q, first generates a reasoning trace o_{think} (within `<think>`/`</think>` tags) and then a final answer o_{ans} (within `<answer>`/`</answer>` tags).

Multiple-choice QA is widely used in multimodal training and evaluation: it supports binary rewards without external graders, making it ideal for GRPO-style post-training. However, models may learn option priors or lexical cues under outcome-only rewards, leading to reasoning–answer inconsistency. To systematically evaluate this, we design two complementary tests: the **CoT Answer Consistency Test** and the **Option Shuffle Consistency Test**.

3.1 CoT Answer Consistency Test

A reasoning MLLM should produce a final answer that is logically consistent with its chain-of-thought (CoT); otherwise, the reasoning cannot be deemed reliable. We empirically observe that GRPO training increases the frequency of contradictions between CoT and final answer. To quantify this, we adopt an LLM-as-Judge protocol. Let f_{judge} denote the judge model and P_{judge} its evaluation prompt. For each example $d \in D_{\text{test}}$, the model outputs a reasoning trace o_{think} and a final answer o_{ans}. The *CoT and Answer Consistency Rate* (CACR) is defined as

$$\text{CACR} = \frac{1}{|D_{\text{test}}|} \sum_{d \in D_{\text{test}}} \mathbf{1}\left[f_{\text{judge}}(d) = \text{“consistent”}\right], \tag{1}$$

where $f_{\text{judge}}(d) = f_{\text{judge}}\left(P_{\text{judge}}, o_{\text{think}}^{(d)}, o_{\text{ans}}^{(d)}\right)$, where $\mathbf{1}[\cdot]$ is the indicator function that returns 1 if the judge deems the final answer consistent with the provided reasoning trace, and 0 otherwise. Please refer to the Appendix for the exact specification of P_{judge}.

3.2 Option Shuffling Consistency Test

A robust reasoning MLLM should yield the same final answer when answer options are shuffled, provided that the multimodal evidence and the model's generated reasoning trace are held fixed. Let $\mathcal{S}(\cdot)$ be the option shuffling function applied to the query. The *Option Shuffling Consistency Rate* (OSCR)

(a) CoT and Answer Consistency Rate (CACR) (%)

Model	MathVista	MV-Human	MMVU	MMVU-Human
Qwen2.5-VL-7B-Instruct	77.2	78.0	78.5	74.0
Qwen2.5-VL-7B-CoT-SFT	85.2	88.0	82.3	86.0
Qwen2.5-VL-7B-CoT-SFT-GRPO	81.3	82.0	79.7	82.0
ACRE	**87.5**	**90.0**	**82.8**	**86.0**

(b) Option Shuffling Consistency Rate (OSCR) (%). †Closed-source model included for reference.

Model	TempCompass	MMVU	VideoMME
Gemini-2.5-Pro†	56.0	89.0	77.5
Qwen2.5-VL-7B-Instruct	88.8	97.0	86.4
Qwen2.5-VL-7B-CoT-SFT	5.9	31.4	12.5
Qwen2.5-VL-7B-CoT-SFT-GRPO	10.4	49.8	25.6
ACRE	**17.3**	**74.6**	**29.6**

over a test set D_{test} is defined as

$$\text{OSCR} = \frac{1}{|D_{\text{test}}|} \sum_{d \in D_{\text{test}}} \mathbf{1}\Big[M_r(d) = M_r^{\mathcal{S}}(d)\Big], \tag{2}$$

where $M_r(d) = M_r\big(x^{(d)}, q^{(d)}, o_{\text{think}}^{(d)}\big)$ and $M_r^{\mathcal{S}}(d) = M_r\big(x^{(d)}, \mathcal{S}(q^{(d)}), o_{\text{think}}^{(d)}\big)$, where $\mathbf{1}[\cdot]$ is the indicator function that returns 1 if the two answers are identical and 0 otherwise. Higher values indicate better option shuffling consistency.

3.3 Results Analysis

We follow the inference configuration of Qwen2.5-VL (detailed in Sec. 5). Table 1a reports CACR on MathVista (Lu et al., 2023a) and MMVU (Zhao et al., 2025); Table 1b reports OSCR on three video reasoning benchmarks. To validate the LLM-as-judge results, we additionally include human expert evaluations on 50-sample subsets of MathVista and MMVU (denoted MathVista-Human and MMVU-Human).

CACR patterns across training regimes. CoT-SFT strengthens CACR while GRPO erodes it. Moving from instruction-only to CoT-SFT yields large CACR gains (MathVista: $77.2 \rightarrow 85.2$; MMVU: $78.5 \rightarrow 82.3$), indicating that supervised exposure to step-by-step traces improves measured consistency between the generated CoT and final answer. GRPO then lowers CACR relative to CoT-SFT across all columns (MathVista: $85.2 \rightarrow 81.3$; MMVU: $82.3 \rightarrow 79.7$), confirming that optimizing for correctness alone can decouple the decision head from the produced trace.

Human–LLM judge agreement. We include human expert evaluations on 50-sample subsets (MathVista-Human and MMVU-Human in Table 1a) to validate the LLM-as-judge protocol. Across all four models, the human CACR scores closely track the automated scores (e.g., MathVista: 78.0 vs. 77.2 for Instruct, 90.0 vs. 87.5 for ACRE), supporting GPT-4o-mini's use as a proxy for consistency judgment in this analysis.

OSCR patterns across training regimes. The Instruct baseline attains the highest OSCR among open-source models (e.g., 97.0 on MMVU), indicating that its decisions are tied to option *content* rather than index tokens. CoT-SFT collapses OSCR dramatically ($88.8 \rightarrow 5.9$ on TempCompass), revealing a strong index-binding effect: supervised traces often conclude with "*thus the answer is (C)*," which becomes brittle under permutation. Vanilla GRPO partially recovers robustness (e.g., $5.9 \rightarrow 10.4$ on TempCompass, $12.5 \rightarrow 25.6$ on VideoMME), but substantial order sensitivity remains.

Summary. GRPO erodes *internal* consistency (CACR) while only partially alleviating *external* order sensitivity (OSCR), indicating that post-trained models lose the ability to reliably retrieve the correct answer from a correct reasoning trace. This motivates the consistency-aware reward design in Sec. 4.

4 Answer-Consistent REinforcement Learning (ACRE)

While GRPO (Guo et al., 2025; Shao et al., 2024) effectively improves outcome accuracy, it can decouple the reasoning trace from the final answer in multimodal multiple-choice settings (Sec. 3): granting a positive reward to WR–CA cases and a negative reward to CR–WA cases is undesirable. To address this, we propose **Answer-Consistent REinforcement Learning** (ACRE), a GRPO-compatible reward shaping scheme that explicitly aligns the generated chain-of-thought with the final answer (Fig. 2).

Consistency check via query option shuffling. Given multimodal input x (e.g., videos or images), a multiple-choice question q with option set $\mathcal{O}$, and ground-truth answer $y \in \mathcal{O}$, the policy first produces a reasoning trace t and an initial answer a:

$$t,\, a \sim \pi_\theta(\cdot \mid x, q).$$

We then apply an option shuffling function $\mathcal{S}(\cdot)$ to obtain a permuted question $\mathcal{S}(q)$ while holding both x and t fixed, and re-prompt the model to

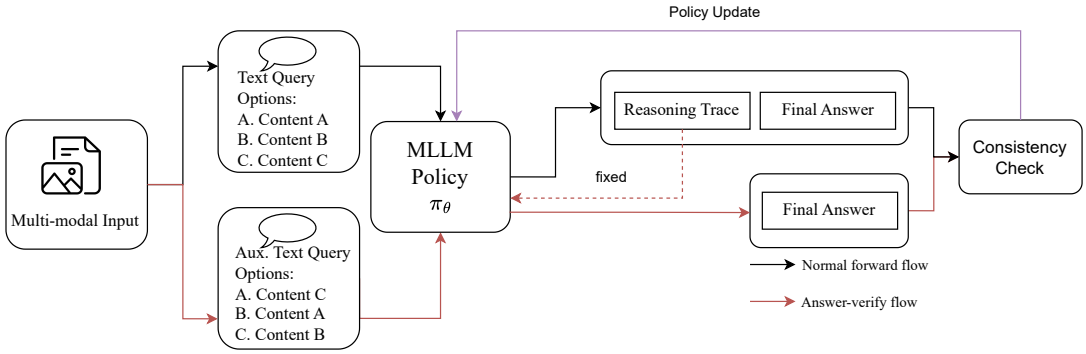


Figure 2: Overview of our proposed ACRE. Given a multimodal input, the MLLM first generates a reasoning trace and a final answer (top path). We then feed the same reasoning trace back to the MLLM along with an auxiliary query where the answer options are shuffled (bottom path). Agreement between the original and shuffled answers is computed after mapping each emitted answer back to its original option identity, and the resulting signal shapes reinforcement learning toward option-shuffling robustness.

produce a second answer $\tilde{a}$:

$$\tilde{a} \sim \pi_\theta(\cdot \mid x, \mathcal{S}(q), t).$$

The auxiliary answer is used only to verify and score the original trajectory; it is not appended to the model output and introduces no extra decoding path at inference time. Agreement is computed over option identity rather than surface letters. Let $\mathrm{id}_q(\cdot)$ map an emitted answer under query q to the original option identity/content it denotes, and let $\mathrm{id}(y)$ be the identity of the gold option. For the shuffled query, $\mathrm{id}_{\mathcal{S}(q)}(\tilde{a})$ maps the emitted shuffled letter back to the corresponding original option. We define

$$\begin{aligned} \mathrm{agree} &= \mathbf{1}\left[\mathrm{id}_q(a) = \mathrm{id}_{\mathcal{S}(q)}(\tilde{a})\right], \\ \mathrm{corr} &= \mathbf{1}\left[\mathrm{id}_q(a) = \mathrm{id}(y)\right], \\ \widetilde{\mathrm{corr}} &= \mathbf{1}\left[\mathrm{id}_{\mathcal{S}(q)}(\tilde{a}) = \mathrm{id}(y)\right]. \end{aligned}$$

Thus, if the original answer is "B" and shuffling moves the same semantic option to "D", the two answers agree only after both are mapped back to the same original option identity.

Consistency-verification reward. Let r_i denote the base reward for trajectory i, which includes outcome correctness and format constraints following DeepSeek-R1 (Guo et al., 2025). ACRE adds a consistency term r_c:

$$r_c = \begin{cases} \alpha_1, & \text{if agree} = 1 \text{ and } \mathrm{corr} = \widetilde{\mathrm{corr}} = 1, \\ \alpha_2, & \text{if agree} = 0 \text{ and } \left(\mathrm{corr} + \widetilde{\mathrm{corr}} = 1\right), \\ \alpha_3, & \text{if agree} = 1 \text{ and } \mathrm{corr} = \widetilde{\mathrm{corr}} = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The four cases capture distinct training signals. *Case 1* (r_c=α_1): the reasoning trace is both correct and shuffle-invariant—the ideal trajectory that should receive the highest reward. *Case 2* (r_c=α_2): the two answers disagree yet exactly one is correct, indicating that the reasoning trace is partially informative but insufficiently grounded to survive option permutation; partial credit encourages the model to tighten the reasoning–answer link without fully discarding useful traces. *Case 3* (r_c=α_3): both answers are wrong yet agree, reflecting internally consistent but factually incorrect reasoning; a small positive reward stabilizes training by discouraging random guessing without rewarding wrong answers. *Case 4* (r_c=0): all remaining configurations—including disagreeing wrong answers—receive no consistency bonus, as neither correctness nor internal coherence is present. We ablate these coefficients in Sec. 5. The final per-trajectory reward is

$$R_i = r_i + r_c. \tag{4}$$

This shaping grants a high bonus only when the reasoning trace yields *stable and correct* answers under option shuffling, while discouraging reliance on spurious patterns such as option-order biases or reasoning–answer drift.

Group-normalized advantages and policy update. As in GRPO, we draw G trajectories $\{o_i\}_{i=1}^G$ per prompt, compute rewards $\{R_i\}$ using equation 4, and normalize within the group:

$$A_i = \frac{R_i - \mathrm{mean}(\{R_j\}_{j=1}^G)}{\mathrm{std}(\{R_j\}_{j=1}^G) + \varepsilon}. \tag{5}$$

The training objective augments GRPO with the consistency-shaped advantages:

$$\mathcal{J}_{ACRE}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(L_i^{\text{clip}} - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right)\right], \tag{6}$$

where $L_i^{\text{clip}} = \min\left(\rho_i(\theta)A_i,\ \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)A_i\right)$ with probability ratio $\rho_i(\theta) = \frac{\pi_\theta(o_i|x,q)}{\pi_{\theta_{\text{old}}}(o_i|x,q)}$, where ϵ is the PPO clipping parameter, β scales the KL regularization to the reference model π_{ref}, and ε is a small constant for numerical stability.

5 Experiments

5.1 Setup

Dataset Construction. We mix Video-QA and Image-QA data for RL training. For Video-QA, we use Open-R1-Video-4.6k (Wang and Peng, 2025). For Image-QA, we sample a 4.6k subset of multiple-choice questions (Math, Chart, OCR, Knowledge, and Spatial) from Video-R1-260k (Feng et al., 2025). Together these form ACRE-9.2k.

Implementation details. We adopt Qwen2.5-VL-7B (Bai et al., 2025) as the base MLLM. Following DeepSeek-R1, training proceeds in two stages: SFT cold start followed by RL. We use the SFT model provided by Video-R1 (Feng et al., 2025), trained on Video-R1-CoT-165k (chain-of-thought annotated image and video samples), denoted Qwen2.5-VL-7B-CoT-SFT. In the second stage, we train the CoT-SFT model on ACRE-9.2k using GRPO and ACRE, yielding Qwen2.5-VL-7B-CoT-SFT-GRPO and **ACRE**, respectively. The consistency-shaping coefficients are $(\alpha_1, \alpha_2, \alpha_3)$=(1, 0.9, 0.3). We also adopt the length-based reward from Video-R1 to regulate output length. Specifically, this mechanism aims to strike a balance between encouraging deeper reasoning and preventing overthinking. For each reasoning path o_i, if the predicted answer is correct and the response length falls within a predefined interval $[l_{\min}, l_{\max}]$, the model receives an additional reward $r_l = \omega$. Formally:

$$R_i' = \begin{cases} R_i + \omega, & \text{if } o_i \text{ is correct and} \\ & l_{\min} \le \text{len}(o_i) \le l_{\max}, \\ R_i, & \text{otherwise}, \end{cases} \tag{7}$$

where R_i' is the final reward after incorporating the length bonus. The hyperparameters are set to $\omega = 0.2$, $l_{\min} = 320$, and $l_{\max} = 512$.

For the CoT Answer Consistency Test, we use GPT-4o-mini as f_{judge}. For the Option Shuffling Consistency Test, we implement $\mathcal{S}$ as follows: for a multiple-choice query $q = q_t + q_o$ (text query q_t and option set q_o), $\mathcal{S}(q) = q_t + \text{derangement}(q_o)$, where derangement is a permutation in which no element appears in its original position.

Evaluation Datasets. We evaluate our model on three video benchmarks and two multimodal math benchmarks: VideoMME (Fu et al., 2025), MMVU (Zhao et al., 2025), TempCompass (Liu et al., 2024), MathVerse (Zhang et al., 2024), and MathVista (Lu et al., 2023a). For MathVerse and MathVista, evaluations are performed on their corresponding multiple-choice QA subset. For all evaluations, we follow the decoding configuration used in the official Qwen2.5-VL demo, with top_p = 0.001 and temperature = 0.01.

5.2 Main Results

Overall Performance. As shown in Table 2, results across five benchmarks validate the effectiveness and *data efficiency* of ACRE for both video reasoning and multimodal math reasoning. Starting from the same CoT-SFT initialization, replacing vanilla GRPO with ACRE yields +2.2 points on the *Video Reasoning Avg.* (0.615 → 0.637) and +1.5 points on the *Math Avg.* (0.583 → 0.598), demonstrating that enforcing reasoning–answer agreement improves both accuracy and decision robustness. We further assess *generalization*: both methods are RL-finetuned *only* on video QA data (OpenR1-Video-4.6k) with **no** math QA exposure. Table 3a shows that ACRE surpasses GRPO on both MathVista (68.8 vs. 67.3) and MathVerse (45.7 vs. 44.5), yielding +1.5 and +1.2 point gains. Compared with Video-R1-7B, whose RL stage uses Video-R1-260k (28× more RL-stage data than ACRE-9.2k), ACRE remains competitive (+0.4 on Video Reasoning, +1.6 on Math Reasoning).

ACRE improves measured consistency and option-shuffling robustness. A natural question is whether the gains are limited to accuracy or also appear in reasoning–answer consistency and option perturbation tests. Tables 1a and 1b show improvements on both metrics. First, CACR measures whether the CoT is judged consistent with the final answer, not just whether the answer is correct. ACRE achieves MathVista CACR of **87.5** (+2.3 vs. CoT-SFT, +6.2 vs. GRPO) and MMVU CACR of **82.8** (+0.5 vs. CoT-SFT, +3.1 vs. GRPO), match-

Table 2: Accuracy comparison on video reasoning and multimodal math benchmarks. The Train-data column indicates the data used in the RL phase; GRPO and ACRE use the same CoT-SFT initialization and RL data unless otherwise noted.

Model	Train-data	Video Reasoning				Math		
		VideoMME	MMVU	TempCompass	Avg.	MathVerse	MathVista	Avg.
Qwen2.5-VL-7B-Instruct	-	0.503	0.598	0.708	0.603	0.421	0.685	0.553
Qwen2.5-VL-7B-CoT-SFT	-	0.528	0.613	0.692	0.611	0.456	0.680	0.568
Qwen2.5-VL-7B-CoT-SFT-GRPO	ACRE-9.2k	0.542	0.608	0.695	0.615	0.461	0.704	0.583
Video-R1-7B	Video-R1-260k	**0.558**	0.629	**0.713**	0.633	0.477	0.687	0.582
ACRE	ACRE-9.2k	0.545	**0.656**	0.710	**0.637**	**0.481**	**0.715**	**0.598**

ing or exceeding SFT-level CACR on the evaluated settings. Second, the OSCR improvement (e.g., 74.6% on MMVU versus 49.8% for GRPO) shows stronger robustness when the same trace is evaluated under shuffled option surfaces. Importantly, the accuracy and consistency gains are complementary: ACRE's higher accuracy is consistent with reducing CR–WA-style failures where correct reasoning was previously followed by wrong answer selection.

Visualizations. Fig. 3 contrasts attention allocation for ACRE (left) and GRPO (right) when generating the final answer token on a Video QA example (ground truth: **A**). We group output reasoning tokens into blocks of n=50 and compute average attention per group. Under GRPO, attention collapses onto *system prompt* tokens (≈ 0.02), while content-bearing regions—question, text query, and options—receive weak weights. Despite option A receiving the highest average attention among options, the GRPO model incorrectly predicts **C**, symptomatic of index/format shortcuts. In contrast, ACRE redistributes attention toward *options* and *output tokens*, with a smoother spread across option spans and a pronounced rise near the final output groups. This provides qualitative evidence consistent with the model attending more to option *content* than positional cues, in line with the higher CACR and OSCR. Additional qualitative comparisons are shown in Fig. 4.

5.3 Ablations

Hyperparameters. We ablate the consistency-shaping coefficients $\alpha_1, \alpha_2, \alpha_3$ (Sec. 4). Fixing α_1=1, we vary $\alpha_2 \in \{1.0, 0.9, 0.8, 0.7\}$ with α_3=0.3, obtaining MMVU scores of 0.645, **0.656**, 0.637, 0.638. The optimal α_2=0.9 indicates that disagreeing-but-one-correct cases should receive reward close to—but strictly below—α_1; both over-rewarding (α_2=1.0) and under-rewarding ($\alpha_2 \leq 0.8$) degrade performance. Fixing α_2=0.9, we vary $\alpha_3 \in \{0.0, 0.3, 0.5\}$, obtaining 0.643, **0.656**, 0.629. A moderate α_3=0.3 stabilizes learning by encouraging internally consistent traces, while no shaping (α_3=0.0) or excessive shaping (α_3=0.5) both hurt. The best setting is $(\alpha_1, \alpha_2, \alpha_3) = (1, 0.9, 0.3)$. Overall, the ablation confirms that all three consistency-shaping terms contribute: removing any one degrades performance, and the reward hierarchy $\alpha_1 > \alpha_2 > \alpha_3 > 0$ is essential for stable training.

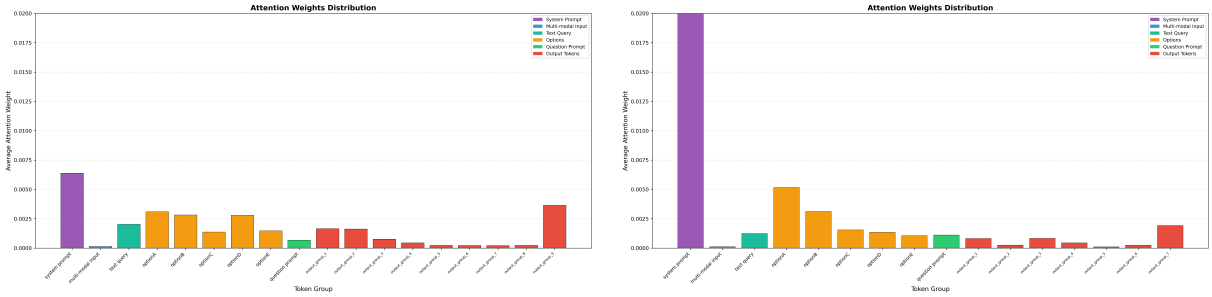
Training overhead. The consistency reward requires a second forward pass per trajectory. On ACRE-9.2k, total GPU hours increase from 8.2 (GRPO) to 10.1 (ACRE), a +23% overhead. This is modest because the second pass only decodes the final answer token given the cached reasoning trace, not the full CoT. Further reduction is possible by reusing the KV cache from the first pass.

5.4 Beyond Multiple-Choice: Open-Ended Generalization

While ACRE's consistency reward is defined over multiple-choice options, its core principle—that a reasoning trace should yield a stable answer under surface-level perturbations—may extend beyond MCQ settings. We evaluate two strategies to extend ACRE beyond MCQ settings.

Mixed-reward strategy. Real-world training data mixes formats. ACRE supports a flexible reward mechanism: the consistency reward r_c is applied to MCQ samples, while standard outcome-based or reference-based rewards (e.g., correctness verification, ROUGE) handle open-ended tasks. This setup tests whether consistency shaping can coexist with standard open-ended rewards and may transfer some benefit beyond option selection.

We validate this with **ACRE-6k**, a 1:1 mixture of MCQs and open-ended QAs from Video-R1-260k (Feng et al., 2025). As shown in Ta-



(a) Attention Visualization of ACRE

(b) Attention Visualization of GRPO

Figure 3: Attention visualization comparison between ACRE and GRPO when generating the final answer token. ACRE redistributes attention toward option content and output tokens, while GRPO concentrates on system prompt tokens.

(a) Generalization to math reasoning when trained only on video QA data.

Model	Train data	MathVista	MathVerse
GRPO	OpenR1-Video-4.6k	0.673	0.445
ACRE	OpenR1-Video-4.6k	**0.688**	**0.457**

(b) Impact of mixed-reward strategy on ACRE-6k (1:1 MCQ + open-ended QA).

Method	Train Data	MMVU (MCQ, Acc %)	NextQA (Open, WUPS)
GRPO (Baseline)	ACRE-6k	61.8	19.78
ACRE (Ours)	ACRE-6k	**63.0**	**21.97**
Improvement	-	**+1.2**	**+2.19**

ble 3b, ACRE trained on mixed data outperforms the GRPO baseline on both MMVU (MCQ, +1.2) and the NextQA open-ended split (+2.19 WUPS, 21.97 vs. 19.78). The open-ended gain is notable because NextQA requires free-form answers evaluated by WUPS similarity—not option selection—suggesting that consistency shaping can coexist with standard open-ended rewards and may transfer some benefit. Broader open-ended evaluation is needed to establish the scope of this transfer.

Automated conversion via distractor generation. When only open-ended data is available, we synthesize consistency signals through automated distractor generation. Given an instruction tuning sample $d{=}(x, q, r)$, we prompt GPT-4o-mini to generate three hard-negative candidates (r_1, r_2, r_3) that are topically relevant but factually flawed (e.g., altered actions, objects, or attributes). The query is reformatted into a selection task q', yielding a synthetic MCQ tuple $d'{=}(x, q', \{r, r_1, r_2, r_3\})$ for ACRE training. We provide the distractor generation prompt and preliminary results in Appendix A.4.

6 Conclusion

We studied reasoning–answer inconsistency in outcome-only RL for multimodal multiple-choice reasoning. Using two complementary metrics—CACR and OSCR—we showed that vanilla GRPO improves accuracy yet erodes measured consistency between the chain-of-thought and the final answer. To address this, we introduced ACRE, a GRPO-compatible reward shaping scheme that enforces shuffle-invariant agreement conditioned on correctness. Across five video and math reasoning benchmarks, ACRE yields consistent gains over GRPO (+2.2 and +1.5 average points for video and math, respectively) while improving measured CoT–answer consistency and option-shuffling robustness. These results indicate that coupling outcome optimization with explicit consistency verification yields models with stronger measured reasoning–answer consistency and more robust decisions under option perturbations.

Future work. Our framework is agnostic to the specific perturbation: option shuffling can be replaced with LLM-generated paraphrases or back-translation for broader consistency enforcement. Combining ACRE with process reward models that supervise intermediate steps is another promising direction. We also plan to investigate whether ACRE's gains persist or amplify at larger model scales.

Limitations

ACRE's consistency reward is most directly applicable to multiple-choice settings where option shuffling is well-defined. While our mixed-reward experiments (Sec. 5.4) provide preliminary evidence of transfer to open-ended QA, extending the con-

sistency signal to fully open-ended generation—where no discrete option set exists—remains an open challenge. Additionally, our experiments use a single base model (Qwen2.5-VL-7B); evaluating ACRE across model scales and architectures would strengthen the generality of our findings.

Ethics Statement

This work improves the reliability and consistency of reasoning in multimodal large language models. By encouraging models to produce reasoning traces logically aligned with their final answers, our work contributes to more trustworthy and interpretable AI systems. The improved consistency could benefit applications requiring explainable AI, such as educational tools and decision support systems.

References

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv 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. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

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*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

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

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Goyal. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Pan Lu, Baolin Peng, Hao Cheng, Michel Gal, Andriy Amini, Amanpreet Singh, Ming-Wei Chang, Miri Geva, Ming-Wei Zhou, Tulsee Rether, and 1 others. 2023b. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Kevin Hall, Luyu Gao, Sarah Wiegreffe, Uri Alon, Peter Andersen, Ankur Misra, Hila Arnaout, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

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

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-salmonn-o1: Reasoning-enhanced audio-visual large language model. *arXiv preprint arXiv:2502.11775*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489.

A Appendix

A.1 LLM Usage Statement

We clarify that the use of LLMs in this study is restricted to writing assistance, specifically for grammar correction and enhancing readability. No LLM was involved in the research design, experimental execution, or data analysis. The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

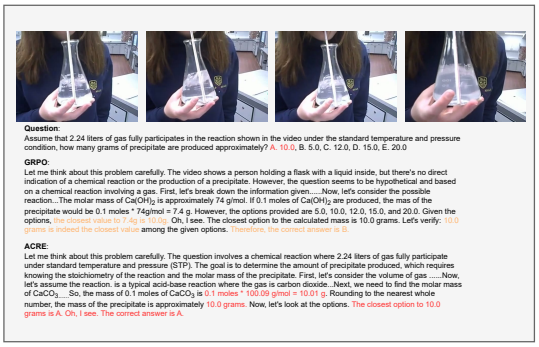


Figure 4: Visualizations of GRPO and ACRE. Red denotes correct answer or reasoning trace and orange denotes flawed answer or reasoning trace.

A.2 Prompts used in Sec. 3

The prompts are included in the supplementary materials as judge_prompt.md

A.3 Codes

Codes are provided in the supplementary materials as ACRE_code.zip

A.4 Automated Distractor Generation for Open-Ended Tasks

To extend ACRE to open-ended data, we convert free-form QA pairs into synthetic MCQs via distractor generation. Given a sample (x, q, r) with ground-truth answer r, we prompt GPT-4o-mini with the following template:

Distractor Generation Prompt

Given the following question and correct answer about a video/image, generate exactly 3 plausible but incorrect answer options. Each distractor should:
1. Be topically relevant to the question
2. Contain a specific factual error (e.g., wrong action, object, attribute, or quantity)
3. Be similar in length and style to the correct answer

Question: {q}
Correct Answer: {r}
Output exactly 3 distractors, one per line.

The original query is then reformatted as a 4-option MCQ $d' = (x, q', \{r, r_1, r_2, r_3\})$ with the correct answer placed at a random position. This synthetic MCQ can then be used with ACRE's consistency reward during training. We leave a comprehensive evaluation of this conversion pipeline at scale to future work.

A.5 More Visualizations