DyCo-RL: Dynamic Cross-Modal Coordination for Visual Reasoning

Anonymous ACL submission

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a leading paradigm for enhancing visual reasoning in Multimodal Large Language Models (MLLMs). However, existing RLVR methods optimize primarily for the reasoning outcome, fundamentally overlooking the fine-grained cross-modal coordination required during the generation process. Through token-level analyses and controlled interventions, we reveal that during Chain-of-Thought (CoT) reasoning, MLLMs frequently fail to dynamically alternate between extracting visual evidence and synthesizing textual context—a coordination breakdown that is causally linked to reasoning failures. Motivated by these findings, we propose DyCo-RL, which integrates dynamic cross-modal coordination into RLVR optimization. Specifically, DyCo-RL uses the Fisher–Rao geodesic distance to measure within-modality attention shifts, assigning tokens to either visually-oriented or text-oriented functional roles. It then evaluates the alignment between a token's actual attention allocation and its assigned role, leveraging this score for alignment-guided advantage reweighting during policy optimization. Extensive experiments demonstrate that the algorithm-agnostic DyCo-RL, when applied to Qwen2.5-VL-3B/7B, consistently improves four representative RLVR algorithms across seven benchmarks spanning visual-centric and mathematical reasoning.

1 Introduction

Visual reasoning (Ke et al., 2025; He et al., 2021) requires progressively acquiring, interpreting, and integrating information across visual and textual modalities. Reinforcement Learning with Verifiable Rewards (RLVR) (Shao et al., 2024; Zheng et al., 2025; Gao et al., 2025; Yu et al., 2025), originally developed for LLMs, has recently been extended to MLLMs as a unified framework that jointly improves visual perception and multi-step reasoning through end-to-end reward-driven training (Wang et al., 2025d; Huang et al., 2025; Rezaei et al., 2026; Zhang et al., 2025a).

However, visual reasoning is inherently a dynamic process. As the Chain-of-Thought (CoT) unfolds, the model must continuously alternate between faithfully extracting visual evidence at certain steps and strictly anchoring its logic to the preceding textual context at others. When this cross-modal coordination breaks down, distinct failure modes emerge. As illustrated in Fig. 1(a), the model first hallucinates visual features by mistakenly asserting that angle ADE is 80°, and subsequently suffers from logical incoherence by baselessly claiming that angle CED equals angle ADE without valid justification from its own reasoning history. Despite recent advances, existing RLVR methods leave this dynamic coordination problem unaddressed. They either apply uniform policy optimization across all tokens without distinction (Shao et al., 2024; Zheng et al., 2025), or focus exclusively on strengthening visual perception (Yang et al., 2025; Wang et al., 2025d; Huang et al., 2025), thereby neglecting the intricate interplay required for robust multi-step reasoning.

To uncover the underlying mechanism behind these failures, we investigate the rationale generation process through the lens of token-level attention dynamics—specifically, how a token's attention weights are distributed between visual patches and textual context (Zhang et al., 2025b; Hassanin et al., 2024; Guo et al., 2022). As illustrated in Fig. 1(b), we first observe a *correlation*: in erroneous samples, visually-oriented tokens systematically under-attend to the image, while text-oriented tokens fail to sufficiently anchor to the preceding textual context. Crucially, we further establish *causality* through controlled intervention experiments. We demonstrate that dynamically amplifying the appropriate modality attention based on a token's functional role consistently rectifies these

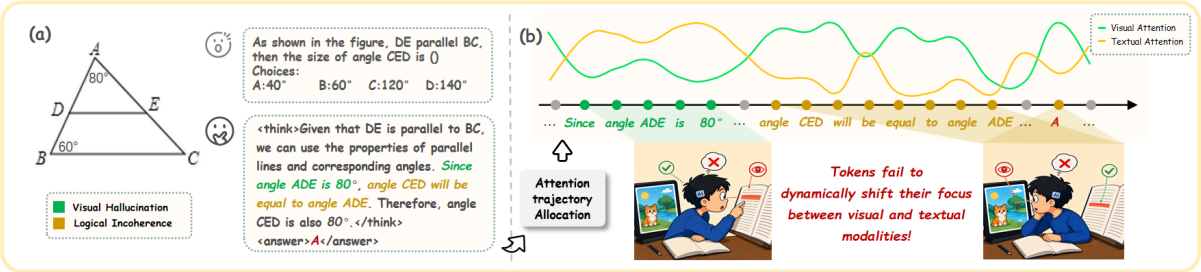


Figure 1: An illustrative example of reasoning failures in Qwen2.5-VL-3B optimized via GRPO on ThinkLite-hard-11K (Wang et al., 2025b). (**a**) The model first hallucinates visual features (mistakenly asserting $\angle\text{ADE} = 80°$), and subsequently generates logically incoherent text (baselessly claiming $\angle\text{CED} = \angle\text{ADE}$, contradicting its own reasoning history). (**b**) Token-level attention trajectory analysis reveals the root cause: visually-oriented tokens under-attend to the image, while text-oriented tokens fail to sufficiently anchor to the preceding textual context.

reasoning failures and recovers the correct answers.

Building on these findings, we propose **Dy**namic **Co**ordination **R**einforcement **L**earning (**DyCo-RL**) to integrate token-level cross-modal coordination directly into the RLVR framework. DyCo-RL operates in two stages. First, by computing the Fisher–Rao geodesic distance between consecutive within-modality attention distributions, it assigns each token a functional role (e.g., visually- or text-oriented); a larger distance reflects significant attention restructuring, indicating active information extraction from that modality. Second, it evaluates the alignment between the token's actual attention allocation and its assigned role. This alignment score is then leveraged for alignment-guided advantage reweighting during policy optimization, effectively amplifying the learning signals for well-coordinated tokens while attenuating the influence of misaligned ones. Designed as an algorithm-agnostic plug-in, DyCo-RL can be seamlessly incorporated into mainstream RL algorithms. In summary, our main contributions are as follows:

- We identify cross-modal coordination breakdowns as a critical bottleneck in visual reasoning for MLLMs. Through token-level correlational analysis and causal intervention experiments, we reveal that tokens frequently fail to dynamically shift their focus between modalities during CoT generation.
- We propose DyCo-RL, a novel approach that embeds dynamic, token-level cross-modal coordination into the RLVR framework. By leveraging the Fisher–Rao geodesic distance of attention shifts to assign functional roles, DyCo-RL dynamically reweights token-level advantages based on modality-specific alignment signals.
- We demonstrate that DyCo-RL is a highly versatile plug-in. Extensive evaluations on Qwen2.5-VL-3B/7B show that it consistently improves four representative RLVR algorithms (GRPO, DAPO, SAPO, and GSPO) across seven diverse benchmarks spanning both visual-centric and mathematical reasoning. Crucially, ablation studies confirm that these gains stem from improved dynamic coordination rather than naive modality bias.

2 Related Work

RLVR for Visual Reasoning. Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a dominant paradigm for enhancing visual reasoning in MLLMs. Foundational algorithms such as GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), GSPO (Zheng et al., 2025), and SAPO (Gao et al., 2025) optimize policy gradients primarily at the trajectory level, applying uniform credit assignment across all tokens. To explicitly enhance visual perception, PAPO (Wang et al., 2025d) introduces an implicit perception loss via KL divergence between outputs under original and masked visual inputs, encouraging visually grounded reasoning at the loss level. Progressing toward finer granularity, StepGRPO (Li et al., 2026a) incorporates dense step-level rewards to evaluate the accuracy and logical consistency of intermediate reasoning steps. Finally, at the token level, methods like TokenDPO (Zeng et al., 2024) and Critical Tokens (Lin et al., 2025) have demonstrated the efficacy of fine-grained optimization for preference alignment in LLMs. In the multimodal domain, VPPO (Huang et al., 2025) filters policy gradients to emphasize highly visually-dependent tokens, and AT-RL (Jiao et al., 2026) selectively reinforces anchor tokens exhibiting strong cross-modal connectivity via graph-based attention clus-

tering. Despite these advances, existing token-level MLLM approaches assign credit based on a single-modality prior. They overlook the dynamic nature of CoT reasoning, where different tokens must alternate their focus between visual and textual contexts. Our work bridges this critical gap by jointly coordinating both visual and textual modality demands during token-level policy optimization.

Cross-Modal Attention Analysis in MLLMs. Attention patterns serve as a crucial window into the cross-modal interactions of MLLMs. One line of research leverages attention for inference-time interventions; for instance, attention refocusing and visual-aware decoding methods (Wang et al., 2025a; Zhang et al., 2025b) dynamically reweight attention distributions during generation to mitigate hallucinations and improve visual faithfulness. Another line of research investigates attention from an interpretability perspective. Recent studies (Wang et al., 2025c; Team et al., 2026; Li et al., 2026b; de Llano et al., 2026) demonstrate that distinct attention dynamics correlate strongly with specific functional behaviors, suggesting that the evolution of attention states intrinsically reflects the model's underlying reasoning process. Crucially, however, these approaches utilize attention solely as an inference-time heuristic or a post-hoc diagnostic tool. In contrast, DyCo-RL is the first to fundamentally integrate dynamic attention coordination directly into the RLVR training objective.

3 Preliminaries

3.1 Modality-Specific Attention Decomposition

In autoregressive MLLMs, the language model decoder attends to both visual and textual tokens at each generation step. Let the context be partitioned into a set of visual token indices $\mathcal{I}_{\text{vis}}$ and a set of textual token indices $\mathcal{I}_{\text{txt}}$ (comprising the input prompt and all previously generated tokens). At the last decoder layer, the attention weight from a generated token at position i to a preceding context token at position j, averaged over all attention heads, is computed as:

$$a_{i,j} = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \frac{\exp(e_{ij}^h)}{\sum_k \exp(e_{ik}^h)}, \quad e_{ij}^h = \frac{\mathbf{q}_i^h \cdot (\mathbf{k}_j^h)^\top}{\sqrt{d}} \tag{1}$$

where $\mathbf{q}_i^h$ and $\mathbf{k}_j^h$ are query and key vectors for head h, and d is the head dimension. For a generated token at step t, we define its *modality-specific attention score* to modality $m \in \{\text{vis}, \text{txt}\}$ as:

$$r_t^m = \sum_{j \in \mathcal{I}_m} a_{t,j} \tag{2}$$

which quantifies the total attention density allocated to each modality during generation.

3.2 Group Relative Policy Optimization (GRPO)

Given an input prompt x, the reference policy $\pi_{\theta_{\text{old}}}$ generates a group of G candidate responses $\{o_1, \ldots, o_G\}$. Under the RLVR framework, each response receives an outcome-based reward R_i determined by whether its final answer matches the ground truth. To reduce variance and eliminate the need for a separate value network, GRPO computes a group-normalized advantage for each response:

$$\hat{A}_i = \frac{R_i - \mu(R_{1:G})}{\sigma(R_{1:G})} \tag{3}$$

where $\mu(R_{1:G})$ and $\sigma(R_{1:G})$ are the mean and standard deviation of the rewards within the group. The policy π_θ is then optimized via a clipped surrogate objective, where the scalar advantage $\hat{A}_i$ is uniformly applied across all token positions within a response:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(\rho_{i,t}\,\hat{A}_i, \\ \text{clip}\big(\rho_{i,t}, 1-\epsilon, 1+\epsilon\big)\hat{A}_i\Big)\Bigg] \tag{4}$$

where $\rho_{i,t} = \pi_\theta(o_{i,t}|x, o_{i,<t}) \,/\, \pi_{\theta_{\text{old}}}(o_{i,t}|x, o_{i,<t})$ denotes the importance sampling ratio at step t.

This formulation introduces a fundamental limitation for MLLMs: a single scalar advantage is broadcast uniformly to all tokens. Consequently, it provides identical learning signals to tokens whether they function to extract visual evidence or synthesize textual logic, fundamentally ignoring the fine-grained, dynamic modality coordination required for robust visual reasoning.

4 Method

We first analyze the cause of cross-modal coordination breakdowns in visual reasoning (Section 4.1). Building on these findings, we propose **Dy**namic **Co**ordination **R**einforcement **L**earning (**DyCo-RL**), which operates in two stages: (i) assigning tokens to visually- or text-oriented roles via

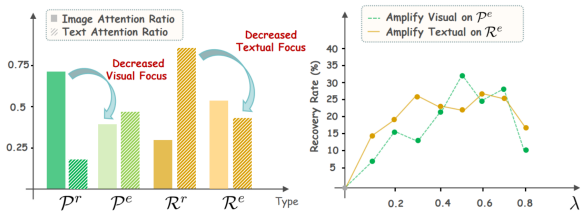


Figure 2: **(Left)** Correlation analysis of modality-specific attention. $\mathcal{P}/\mathcal{R}$ denote visually/text-oriented tokens, and superscripts r/e indicate correct/erroneous states. **(Right)** Causal intervention on erroneous tokens. The green (yellow) curve shows the effect of amplifying visual (textual) attention for $\mathcal{P}^e$ ($\mathcal{R}^e$). Both targeted interventions yield a consistent recovery rate under moderate enhancement.

the Fisher–Rao geodesic distance between consecutive within-modality attention distributions (Section 4.2.1); and (ii) evaluating the alignment between actual attention allocation and assigned roles to perform alignment-guided advantage reweighting during policy optimization (Section 4.2.2).

4.1 Diagnosing Cross-Modal Coordination Breakdowns

We examine cross-modal coordination breakdowns at token-level granularity from two complementary perspectives: **correlation analysis**, which examines whether attention patterns systematically differ between correct and erroneous tokens, and **causal intervention**, which tests whether correcting attention misalignment recovers model performance.

Correlation Analysis. We collect 200 erroneous rollouts from MathVerse (Zhang et al., 2024) and MathVision (Wang et al., 2024) which generated by Qwen2.5-VL-3B optimized via GRPO on ThinkLite-hard-11K. Each generated token span is categorized according to whether it primarily performs image-grounded perception or text-based reasoning within the local decoding context. Specifically, we distinguish *visually-oriented* tokens ($\mathcal{P}$), which extract or describe information from the image, from *text-oriented* tokens ($\mathcal{R}$), which conduct logical inference based on the preceding textual context. Each category is further divided by correctness, yielding four disjoint groups: $\mathcal{P}^r$ (correct), $\mathcal{P}^e$ (erroneous), $\mathcal{R}^r$ (correct), and $\mathcal{R}^e$ (erroneous). Labels are assigned by annotators at the semantic-span level and then projected back to tokens. Full annotation details are provided in Appendix F.

For each group $\mathcal{S} \in \{\mathcal{P}^r, \mathcal{P}^e, \mathcal{R}^r, \mathcal{R}^e\}$, we average the modality-specific attention score r_t^m (Eq. 2) over all tokens in the group. As shown in Fig. 2, correct visually-oriented tokens ($\mathcal{P}^r$) allocate significantly more attention to the image than erroneous ones ($\mathcal{P}^e$). Symmetrically, correct text-oriented tokens ($\mathcal{R}^r$) attend more to the preceding context than erroneous ones ($\mathcal{R}^e$). These results reveal a clear association between modality-specific attention allocation and token-level correctness.

Causal Intervention. To test whether the observed attention misalignment causally contributes to reasoning errors, rather than merely correlating with them, we design a controlled attention enhancement experiment. For tokens in $\mathcal{P}^e$ or $\mathcal{R}^e$, we selectively amplify their attention toward the modality their functional role demands: visual inputs for visually-oriented erroneous tokens ($\mathcal{M} = \mathcal{I}_{\text{vis}}$) and textual context for text-oriented erroneous tokens ($\mathcal{M} = \mathcal{I}_{\text{txt}}$). Specifically, we dynamically modify the aggregated attention weight from the generated token at step t to a preceding context token at position j as follows:

$$\tilde{a}_{t,j} = \frac{a_{t,j} \cdot \left(1 + \lambda \cdot \mathbf{1}_{j \in \mathcal{M}}\right)}{\sum_k a_{t,k} \cdot \left(1 + \lambda \cdot \mathbf{1}_{k \in \mathcal{M}}\right)} \tag{5}$$

where $a_{t,j}$ is the original attention weight averaged across all heads (as defined in Eq. 1), $\mathbf{1}_{j \in \mathcal{M}}$ is the indicator function for the target modality, and $\lambda > 0$ controls the enhancement strength.

As shown in Fig. 2, moderate enhancement consistently recovers performance on previously erroneous samples for both functional roles, with visually-oriented tokens showing higher sensitivity to correction. Excessively large λ, however, degrades performance, indicating that over-amplification disrupts the original attention balance. Together with the correlation analysis, these results provide convergent evidence that cross-modal attention misalignment is a significant contributing factor to visual reasoning errors.

4.2 Dynamic Coordination Reinforcement Learning

Based on the findings above, we propose DyCo-RL, whose pipeline is illustrated in Fig. 3.

4.2.1 Token Role Assignment

To identify each token's functional role, we characterize the transition of within-modality attention between consecutive generation steps. A token actively engaging a modality will exhibit substantial redistribution of attention within that modality compared to the previous step.

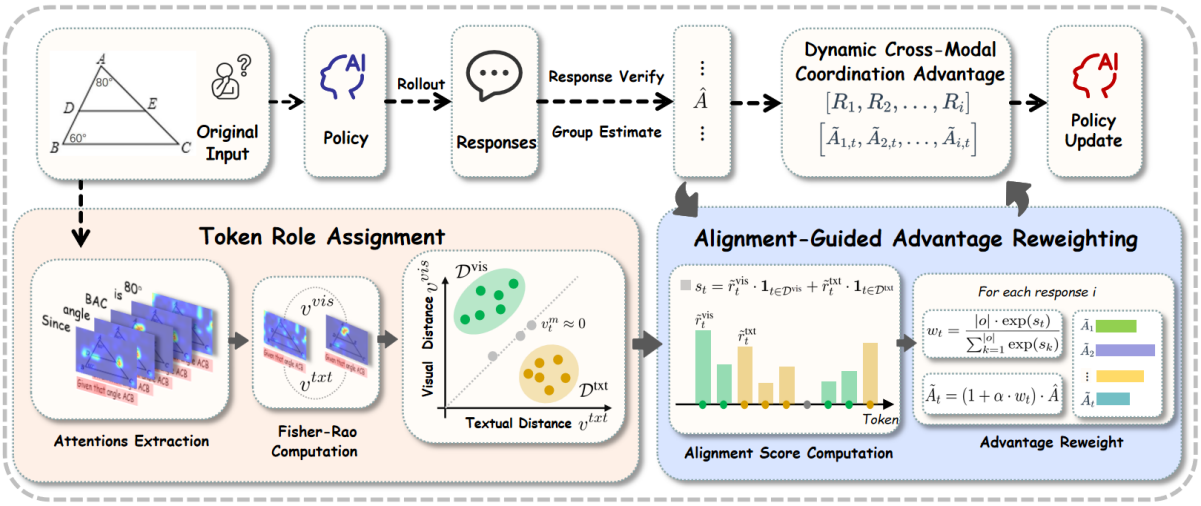


Figure 3: **Overall Pipeline of DyCo-RL.** Instead of directly broadcasting a standard sequence-level advantage ($\hat{A}_i$) to all tokens, our framework transforms it into a fine-grained Dynamic Cross-Modal Coordination Advantage ($\tilde{A}_{i,t}$) for policy updates. Specifically, our dynamic coordination plugin computes the Fisher-Rao attention distance to assign functional roles to individual tokens. By evaluating the alignment between a token's actual attention allocation and its assigned role, we dynamically reweight the standard advantage at the token level.

First, we normalize the attention weights within each modality to obtain a conditional distribution:

$$p^m_{t,j} = \frac{a_{t,j}}{r^m_t}, \quad \forall j \in \mathcal{I}_m \tag{6}$$

which describes how attention is distributed across context tokens within modality m, independent of the total amount allocated to that modality.

Next, we measure the structural change between consecutive within-modality distributions using the Fisher–Rao geodesic distance:

$$v^m_t = 2 \arccos\Big(\sum_{j \in \mathcal{I}_m} \sqrt{p^m_{t-1,\,j}\; p^m_{t,j}} \Big) \tag{7}$$

A high v^m_t indicates substantial structural reallocation of attention within modality m from step $t-1$ to t, reflecting active information acquisition, whereas $v^m_t \approx 0$ indicates inertial behavior. Unlike KL divergence, the Fisher–Rao distance is symmetric and bounded, providing a stable measure for noisy attention dynamics.

Finally, we assign each token to the modality whose attention distribution undergoes stronger structural reconfiguration:

$$\mathcal{D}^m = \{\, t \mid v^m_t - v^{\bar{m}}_t > \tau \,\}, \quad \{m, \bar{m}\} = \{\text{vis}, \text{txt}\} \tag{8}$$

where τ is a stability margin that suppresses ambiguous transitions, and $\bar{m}$ denotes the complementary modality. Tokens falling outside both sets ($\mathcal{D}^{\text{vis}}$ and $\mathcal{D}^{\text{txt}}$) are treated as neutral tokens where neither modality exhibits dominant behavior.

4.2.2 Alignment-Guided Advantage Reweighting

Building on the token role assignment in Section 4.2.1, we translate modality alignment into fine-grained RL optimization guidance.

To derive token-level operations, we temporarily drop the candidate response index i. For the t-th token in a generated sequence, we first measure how well its actual attention matches its assigned functional role using the relatively normalized modality attention score $\tilde{r}^m_t = r^m_t / (r^{vis}_t + r^{txt}_t)$.

$$s_t = \tilde{r}^{\text{vis}}_t \cdot \mathbf{1}_{t \in \mathcal{D}^{\text{vis}}} + \tilde{r}^{\text{txt}}_t \cdot \mathbf{1}_{t \in \mathcal{D}^{\text{txt}}} \tag{9}$$

Consequently, visually- and text-oriented tokens are rewarded for attending to the image and textual context, respectively, while neutral tokens remain unbiased.

Next, instead of broadcasting a single scalar advantage to every token, we reweight it dynamically using the computed alignment score:

$$\tilde{A}_t = (1 + \alpha \cdot w_t) \cdot \hat{A}, \quad w_t = \frac{|o| \cdot \exp(s_t)}{\sum_{k=1}^{|o|} \exp(s_k)}, \tag{10}$$

where $\hat{A}$ is the standard sequence-level advantage, $\alpha > 0$ controls the reweighting strength, and the scaling factor $|o|$ preserves the overall advantage magnitude after softmax normalization.

Finally, generalizing this token-level formulation to the entire candidate group, we integrate the dynamically reweighted advantage $\tilde{A}_{i,t}$ (for the t-th token in the i-th response) into the standard pol-

Table 1: Main Results across four RLVR algorithms and two model scales. Bold formatting indicates the better performance between each baseline algorithm and its DyCo-RL enhanced counterpart.

Model	Mathematical Reasoning			Visual-Centric Reasoning				
	WeMath	MathVision	MathVerse	LogicVista	HallusionBench	MME	MMBench	Avg.
Qwen2.5-VL-3B								
Zero-Shot	20.0	23.0	33.8	34.5	61.3	76.9	62.8	44.6
GRPO	23.0	21.4	34.4	39.8	60.8	79.9	54.8	44.9
GRPO + DyCo-RL	**25.9**	**22.4**	**36.0**	**41.0**	**62.4**	**80.7**	**58.2**	**46.7**
DAPO	**27.4**	17.1	34.0	39.8	60.7	81.0	51.7	44.5
DAPO + DyCo-RL	26.9	**22.8**	**36.6**	**41.4**	**62.7**	**82.1**	**54.8**	**46.8**
SAPO	31.5	**20.1**	35.0	39.4	59.7	**82.1**	53.0	45.8
SAPO + DyCo-RL	**32.0**	19.7	**35.3**	**44.5**	**63.9**	81.1	**54.8**	**47.3**
GSPO	**27.2**	19.7	34.0	35.8	**62.4**	79.6	54.9	44.8
GSPO + DyCo-RL	26.0	**24.0**	**34.1**	**37.4**	62.0	**81.9**	**56.3**	**46.0**
Qwen2.5-VL-7B								
Zero-Shot	33.2	23.7	43.8	44.5	65.0	82.4	82.9	53.6
GRPO	38.7	26.6	47.4	46.2	69.2	83.3	78.0	55.6
GRPO + DyCo-RL	**40.5**	**29.3**	**49.5**	**48.5**	**71.6**	**84.9**	**83.8**	**58.3**
DAPO	36.4	23.0	44.5	42.3	70.0	**86.3**	62.7	52.2
DAPO + DyCo-RL	**37.5**	**24.4**	**45.2**	**43.8**	**71.0**	86.0	**75.8**	**54.8**
SAPO	38.5	**26.6**	47.0	42.3	**71.0**	85.6	74.1	55.0
SAPO + DyCo-RL	**40.7**	24.6	**47.8**	**47.7**	69.2	**86.0**	**74.6**	**55.8**
GSPO	39.4	26.6	42.4	46.5	71.4	81.6	65.8	53.4
GSPO + DyCo-RL	**39.9**	**27.6**	**46.5**	**48.3**	**73.2**	**84.8**	**75.5**	**56.5**

icy gradient objective. Using GRPO as the base algorithm, the final DyCo-RL objective is:

$$\mathcal{L}_{\text{DyCo-RL}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(\rho_{i,t}\,\tilde{A}_{i,t}, \text{clip}\big(\rho_{i,t}, 1-\epsilon, 1+\epsilon\big)\,\tilde{A}_{i,t}\Big)\Bigg] \quad (11)$$

5 Experiments

5.1 Experimental Settings

Models, Data, and Baselines. We integrate DyCo-RL into Qwen2.5-VL-3B and Qwen2.5-VL-7B, utilizing the ThinkLite-hard-11K dataset (Wang et al., 2025b) for training. To demonstrate its versatility, we evaluate DyCo-RL on top of four representative RLVR algorithms spanning diverse optimization paradigms: GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), GSPO (Zheng et al., 2025), and SAPO (Gao et al., 2025). Brief descriptions of each baseline algorithm are provided in Appendix D.

Training Details. Models are trained with a learning rate of 1e-6, a rollout batch size of 4, a maximum response length of 2048 tokens, and a global batch size of 64. For our DyCo-RL plugin, we set the alignment-guided reweighting strength $\alpha = 0.2$ and the token role assignment stability margin $\tau = 0.05$. For fair comparison, we strictly adhere to the official hyperparameters for all RLVR baselines. Comprehensive training configurations are detailed in Appendix E.

Evaluation Benchmarks. We evaluate on seven benchmarks spanning two primary domains: (1) *Mathematical Reasoning*: WeMath (Qiao et al., 2025), MathVision (Wang et al., 2024), and MathVerse (Zhang et al., 2024); and (2) *Visual-Centric Reasoning*: LogicVista (Xu et al., 2025), HallusionBench (Guan et al., 2024), MME (Fu et al., 2023), and MMBench (Liu et al., 2024). We follow the standard evaluation protocol from VLMEvalKit (Duan et al., 2024). All methods are evaluated using top-1 accuracy (Acc@1) at a decoding temperature of 0.1.

5.2 Main Results

Table 1 presents the evaluation results on seven benchmarks across two model scales (Qwen2.5-VL-3B and 7B) and four RLVR algorithms. DyCo-RL consistently enhances baseline performance across the vast majority of benchmarks, confirming that our alignment-guided advantage reweighting serves as a highly effective, plug-and-play module for existing RLVR pipelines.

Algorithm-Agnostic Generality. DyCo-RL yields consistent gains across four structurally diverse optimization algorithms without requiring algorithm-specific hyperparameter tuning. Rather than marginal fluctuations, it improves the average performance across all baselines (GRPO, DAPO, SAPO, and GSPO), with individual benchmark gains reaching up to +5.7. This seamless transferability across clip-based, sequence-level, and soft-gated objectives confirms that DyCo-RL

Table 2: Effectiveness of Fisher–Rao Role Assignment compared to alternative assignment strategies.

Method	WeMath	MathVision	MathVerse	LogicVista	HallusionBench	MME	MMBench	Avg.
GRPO (Baseline)	23.0	21.4	34.4	39.8	60.8	79.9	54.8	44.9
+ Random	23.5	20.1	33.1	36.7	59.7	79.3	54.4	43.8
+ Reverse	24.2	19.4	34.0	36.0	58.6	78.0	52.6	43.3
+ Entropy	25.0	21.0	34.4	40.2	61.2	**81.1**	56.5	45.7
+ KL Divergence	24.7	21.3	35.4	40.6	**62.5**	80.4	57.3	46.1
+ DyCo-RL (Ours)	**25.9**	**22.4**	**36.0**	**41.0**	62.4	80.7	**58.2**	**46.7**

Table 3: Effectiveness of Alignment-Guided Reweighting.

Method	WeMath	MathVision	MathVerse	LogicVista	HallusionBench	MME	MMBench	Avg.
GRPO (Baseline)	23.0	21.4	34.4	39.8	60.8	79.9	54.8	44.9
+ Visual Attn Score ($\tilde{r}_t^{\text{vis}}$)	24.5	**23.0**	32.1	35.4	62.0	78.1	56.3	44.5
+ Text Attn Score ($\tilde{r}_t^{\text{txt}}$)	**26.0**	22.4	35.4	37.8	59.4	78.4	53.0	44.6
+ DyCo-RL (Ours)	25.9	22.4	**36.0**	**41.0**	**62.4**	**80.7**	**58.2**	**46.7**

addresses a fundamental limitation shared by all representative RLVR paradigms.

Consistent Gains Across Domains. The performance boosts span both mathematical and visual-centric reasoning. Crucially, DyCo-RL consistently strengthens evidence grounding on visual tasks (e.g., HallusionBench, MME) while preserving and often enhancing chain-of-thought coherence on complex mathematical benchmarks (e.g., MathVision, MathVerse). This dual improvement firmly supports our central claim: dynamically reweighting advantages based on token roles sharpens visual perception without sacrificing reasoning logic.

Scalability to Larger Models. Beyond algorithmic and task-level generality, the improvements persist on the 7B model, where baselines are inherently stronger and the performance headroom is narrower. DyCo-RL continues to elevate average performance across the board, yielding striking individual surges (e.g., up to +13.1 on MMBench). This robust scaling behavior emphatically indicates that even larger models with advanced capacities still suffer from cross-modal coordination breakdowns, which our method effectively mitigates.

5.3 Ablation Studies

To systematically isolate the performance gains introduced by DyCo-RL, we independently ablate its two core modules: the token role assignment strategy and the alignment-guided reweighting mechanism. All ablation experiments are conducted on the Qwen2.5-VL-3B model using the GRPO baseline. Extended ablation studies and details are detailed in Appendix G.

Effectiveness of Fisher–Rao Role Assignment. We compare DyCo-RL against four alternative token role assignment strategies, structured across progressive levels of complexity (Table 2). First, for basic sanity and directional checks, we evaluate *Random* (arbitrary reweighting) and *Reverse* (inverting our assignment logic). *Random* yields only marginal, unstable gains over GRPO, confirming that unprincipled reweighting merely injects noise. Conversely, *Reverse* causes the most severe degradation across all variants. This inverse penalty provides strong directional proof that our Fisher–Rao metric accurately identifies the correct functional roles.

Next, we examine two competitive heuristic baselines: *Entropy* (a static measure) and *KL* (a dynamic measure). *Entropy* measures the sharpness of the attention distribution at a single step, but it falls short because it inherently conflates predictive uncertainty with cross-modal roles, capturing only a static snapshot. To capture temporal shifts, *KL* uses the KL divergence between consecutive attention distributions. While more competitive, KL divergence is mathematically asymmetric and magnitude-sensitive, failing to capture the true geometric structure of the probability simplex during cross-modal transitions.

Ultimately, the full DyCo-RL method consistently achieves the best performance. By utilizing the Fisher–Rao distance, it provides a symmetric, geometrically principled measure that successfully overcomes the limitations of both static snapshots (Entropy) and asymmetric divergence (KL), correctly aligning each token's learning signal with its true modality role.

Effectiveness of Alignment-Guided Reweighting. To validate the conditional formulation of our alignment score s_t in Eq. 9, we replace it with two modality-biased alternatives while keeping all other components identical. Specifically, rather than dynamically switching the scoring function

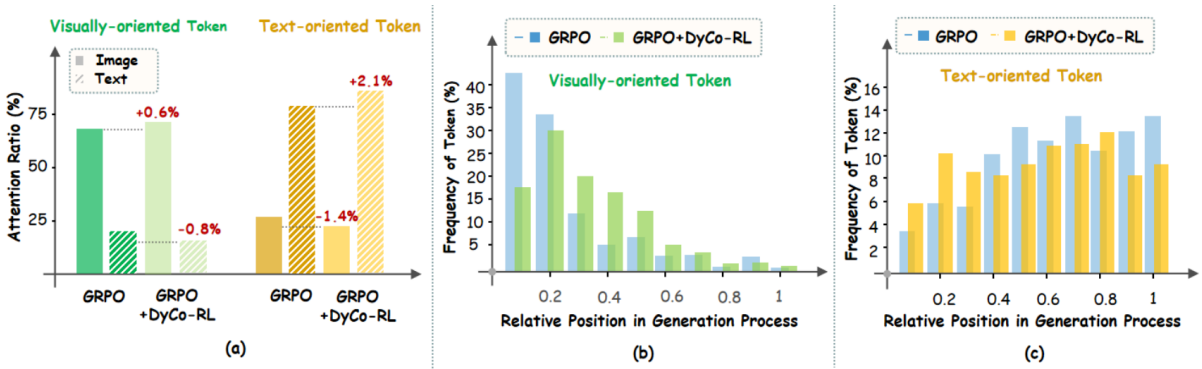


Figure 4: Effectiveness of DyCo-RL in dynamic cross-modal coordination. (a) Alignment between assigned token roles and actual modality attention ratios; (b) Distribution of visually-oriented tokens across the generation process; (c) Distribution of text-oriented tokens across the generation process.

based on a token's assigned role ($\mathcal{D}^{\text{vis}}$ or $\mathcal{D}^{\text{txt}}$), we forcefully apply a single modality's attention ratio to all tokens: *(i) + visual attention score* (setting $s_t = \tilde{r}_t^{\text{vis}}$ universally), and *(ii) + text attention score* (setting $s_t = \tilde{r}_t^{\text{txt}}$ universally). As shown in Table 3, this uni-dimensional reweighting exhibits a clear trade-off: universally up-weighting visual attention improves visually intensive tasks but severely degrades textual reasoning, while the textual variant shows the exact opposite pattern. This confirms that merely encouraging a model to attend to a specific modality is insufficient; coordination is key. In contrast, while single-modality scores occasionally peak on biased tasks, DyCo-RL achieves the most robust and highest average performance across all benchmarks by dynamically aligning each token's advantage with its assigned functional role.

6 Discussion and Analysis

Beyond empirical gains, understanding *why* DyCo-RL succeeds requires a deeper look into the model's internal representations. To this end, we sample 200 generated instances from both the vanilla GRPO baseline and GRPO+DyCo-RL in MathVerse and MathVision. Because the two models generate distinct reasoning trajectories, we independently annotate their tokens into visually-oriented and text-oriented roles following the diagnostic protocol detailed in Section 4.1. Based on this curated evaluation set, we investigate how our alignment-guided reweighting fundamentally reshapes cross-modal reasoning dynamics.

Strengthening Role-Aligned Attention. We first evaluate the internal attention allocations across the annotated tokens. As illustrated in Fig. 4(a), DyCo-RL successfully strengthens the alignment between a token's functional role and its actual modality attention. Specifically, for visually-oriented tokens, the model noticeably increases its attention toward images while suppressing text attention. Conversely, for text-oriented tokens, it actively amplifies text attention at the expense of visual focus. This consistent trend confirms that our advantage reweighting directly encourages tokens to actively extract information from their designated functional modalities.

Reshaping Temporal Dynamics. Furthermore, Fig. 4(b-c) reveals how DyCo-RL reshapes temporal reasoning dynamics. The baseline is rigidly phase-locked: visual perception is front-loaded, while textual reasoning dominates later. DyCo-RL relaxes this boundary. Visually-oriented tokens maintain a sustained presence through the middle stages (e.g., 0.4–0.6) for continuous visual re-grounding, while text-oriented tokens activate earlier and distribute more evenly. This interleaved distribution overcomes the strict "perceive first, reason later" bottleneck, enabling dynamic cross-modal coordination throughout CoT generation.

7 Conclusion

We identify cross-modal coordination breakdowns as a primary bottleneck in visual RLVR, where visually- and text-oriented tokens systematically under-attend to their designated modalities. To address this, we propose DyCo-RL, an algorithm-agnostic plugin that assigns functional token roles via Fisher–Rao attention distance and dynamically reweights advantages based on role-attention alignment. Extensive evaluations show that DyCo-RL consistently improves four representative baselines across seven diverse benchmarks, establishing explicit token-level coordination as an effective path toward faithful multimodal reasoning.

Limitations

Our work has two main limitations. First, although DyCo-RL introduces no additional trainable parameters, computing per-token attention statistics and Fisher–Rao distances during rollout incurs extra training time and GPU memory overhead compared to vanilla RLVR algorithms. Further engineering optimization is needed to reduce this cost, particularly for longer rollouts. Second, due to computational resource constraints, all experiments are conducted on Qwen2.5-VL-3B and 7B models. While we observe consistent improvements across both scales, it remains to be verified whether the proposed mechanism generalizes effectively to substantially larger MLLMs, where attention dynamics may exhibit qualitatively different behavior.

References

Enrique Queipo de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. 2026. Attention sinks and compression valleys in llms are two sides of the same coin. *Preprint*, arXiv:2510.06477.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. 2025. Soft adaptive policy optimization. *Preprint*, arXiv:2511.20347.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. 2022. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3):331–368.

Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad Shahbaz Khan, and Ajmal Mian. 2024. Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, 108:102417.

Feijuan He, Yaxian Wang, Xianglin Miao, and Xia Sun. 2021. Interpretable visual reasoning: A survey. *Image and Vision Computing*, 112:104194.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. 2025. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*.

Zhengbo Jiao, Shaobo Wang, Zifan Zhang, Wei Wang, Bing Zhao, Hu Wei, and Linfeng Zhang. 2026. Credit where it is due: Cross-modality connectivity drives precise reinforcement learning for mllm reasoning. *Preprint*, arXiv:2602.11455.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, Ranjay Krishna, Jiajun Wu, and Hamid Rezatofighi. 2025. Explain before you answer: A survey on compositional visual reasoning. *Preprint*, arXiv:2508.17298.

Weijie Li, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2026a. Step-grpo: Enhancing reasoning quality and efficiency via structured prm-based reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31734–31742.

Xirui Li, Ming Li, and Tianyi Zhou. 2026b. What does rl improve for visual reasoning? a frankenstein-style analysis. *Preprint*, arXiv:2602.12395.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2025. Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 37906–37918. PMLR.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, Yifan Zhang, and 1 others. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.

Ahmad Rezaei, Mohsen Gholami, Saeed Ranjbar Alvar, Kevin Cannons, Mohammad Asiful Hossain, Zhou Weimin, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. 2026. Cppo: Contrastive perception for vision language policy optimization. *arXiv preprint arXiv:XXXX.XXXXX*.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, Ming Wei, Y. Zhang, Fanqing Meng, Chao Hong, Xiaotong Xie, Shaowei Liu, Enzhe Lu, and 18 others. 2026. Attention residuals. *Preprint*, arXiv:2603.15031.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025a. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pages 13712–13736.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025b. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, and Rui Wang. 2025c. Latent space chain-of-embedding enables output-free llm self-evaluation. *Preprint*, arXiv:2410.13640.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and 1 others. 2025d. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. 2025. Look-back:implicit visual refocusing in mllm reasoning.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*.

Chi Zhang, Haibo Qiu, Qiming Zhang, Yufei Xu, Zhixiong Zeng, Siqi Yang, Peng Shi, Lin Ma, and Jing Zhang. 2025a. Perceptual-evidence anchored reinforced learning for multimodal reasoning. *arXiv preprint arXiv:2511.18437*.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025b. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

A Overview of Appendix

- B: **LLM Usage Statement**.
- C: **Broader Impact**.
- D: **Details of Experimental Setup**
- E: **Implementation Details**.
- F: **Diagnostic Annotation Protocol**.
- G: **More Ablation Studies**.
- H: **Analysis of Computational Overhead**.
- I: **Analysis of Training Dynamics**.
- J: **Extended Generalization Analysis**.
- L: **More Visualization Results**.

B LLM Usage Statement

An LLM was used as an assistive tool strictly limited to proofreading and rephrasing for clarity. It was not used for generating core ideas, data analysis, or writing the main content.

C Broader Impact

The core mechanism of DyCo-RL, which leverages within-modality attention restructuring to assign functional roles and perform alignment-guided advantage reweighting, extends naturally beyond visual mathematical reasoning. Cross-modal coordination failures, where a model struggles to dynamically alternate between visual grounding and textual deduction, are a pervasive bottleneck across long-chain multimodal tasks. Whenever an objective requires interleaving perceptual evidence with contextual or symbolic inference, our role-aware token-level optimization becomes directly applicable.

Representative scenarios include visual document and chart understanding, long-video reasoning, embodied agent navigation, and medical image analysis. All these domains require continuous dynamic coordination between perceptual grounding and higher-level reasoning. In such settings, coordination breakdowns frequently manifest as hallucinated details, temporally inconsistent claims, or derivation steps that drift away from the underlying visual evidence.

More broadly, standard RLVR paradigms treat cross-modal coordination as a mere byproduct of sequence-level reward broadcasting. By elevating it to an explicit token-level optimization signal, DyCo-RL provides a robust framework for building faithful, hallucination-resilient multimodal reasoners. Since the Fisher–Rao assignment derives entirely from intrinsic attention geometry and the reweighting strategy remains algorithm-agnostic, this principle integrates seamlessly with existing reward designs and modern policy optimization baselines such as GRPO.

We envision role-aware token-level optimization as a core component for trustworthy multimodal AI systems deployed in high-stakes domains (e.g., healthcare, autonomous systems, and scientific discovery). In these critical fields, the step-by-step fidelity of the reasoning process is just as paramount as the final-answer correctness.

D Details of Experimental Setup

To provide a comprehensive overview of our evaluation framework, we detail the representative baseline algorithms and the selected benchmarks below.

Baseline Algorithms. We evaluate DyCo-RL against four reinforcement learning baselines that span the recent evolution of policy optimization for reasoning-oriented large language models. **GRPO** (Shao et al., 2024) relies on intra-group reward normalization to eliminate the auxiliary critic model. **DAPO** (Yu et al., 2025) extends this by introducing asymmetric clipping and dynamic rollout strategies to stabilize long-chain reasoning. **SAPO** (Gao et al., 2025) utilizes a smooth, sigmoid-based soft gating mechanism with asymmetric temperatures to address optimization discontinuities. Finally, **GSPO** (Zheng et al., 2025) reformulates the objective at the trajectory level, computing a cumulative likelihood ratio across the entire generated sequence for coherent long-chain optimization.

Evaluation Benchmarks. We evaluate DyCo-RL on seven benchmarks that jointly stress the two failure modes our method targets: faithful visual grounding for visually-oriented tokens and consistent textual anchoring for text-oriented tokens. These benchmarks are categorized into mathematical reasoning and visual-centric reasoning domains.

Mathematical Reasoning.

- **WeMath** (Qiao et al., 2025): A hierarchical visual mathematics benchmark that decomposes problem-solving into atomic knowledge

concepts and structured reasoning steps. It provides a diagnostic taxonomy to disentangle knowledge gaps from genuine reasoning failures.

- **MathVision** (Wang et al., 2024): A dataset of 3,040 competition-level math problems with visual contexts, curated from real competitions across 16 disciplines. It is designed to stress long-chain multi-step visual mathematical reasoning.

- **MathVerse** (Zhang et al., 2024): A diagram-centric visual math benchmark where problems are rendered in progressively text-reduced versions. This design isolates the contribution of genuine visual perception by preventing models from relying on textual shortcuts.

- **LogicVista** (Xu et al., 2025): An integrated logical reasoning benchmark requiring models to perform multi-step symbolic inference strictly grounded in visual evidence across spatial, numerical, and mechanical domains.

Visual-centric Reasoning.

- **HallusionBench** (Guan et al., 2024): A diagnostic benchmark explicitly designed to probe visual illusions and language-prior-induced hallucinations in multimodal models through carefully paired image–question controls.

- **MME** (Fu et al., 2023): A comprehensive evaluation suite measuring both perception (e.g., existence, count, position, OCR) and cognition (e.g., commonsense reasoning) across 14 distinct sub-tasks.

- **MMBench** (Liu et al., 2024): An all-around multimodal benchmark comprising multiple-choice questions that span 20 fine-grained ability dimensions, evaluated under a robust circular protocol to mitigate option-order bias.

E Implementation Details

All experiments are conducted within the VLM-R1 training framework (Shen et al., 2025), which serves as the base training paradigm for all models in this work. And as summarized in Table 4, all models are trained for a single epoch on the ThinkLite-VL-Hard-11k dataset, which comprises 11,031 complex reasoning instances. During rollout generation, we sample 4 responses per prompt using a temperature of 0.9. The training objective is guided by a composite reward signal, integrating a binary accuracy reward to verify mathematical correctness and a format reward to enforce structural compliance.

Unlike standard parameter-efficient fine-tuning, we leave the vision tower entirely *unfrozen* to enable full-parameter cross-modal optimization. Across all evaluated baseline algorithms, we employ the AdamW optimizer with a peak learning rate of 1×10^{-6} and a linear decay schedule. To stabilize the training dynamics, we apply gradient checkpointing and cap the global gradient norm at 1.0. For the specific hyperparameter configurations of each RL method (e.g., clipping bounds or KL coefficients), we strictly adhere to their official implementations and recommended settings.

All experiments are executed in `bfloat16` precision, heavily relying on DeepSpeed ZeRO-3 optimization (without CPU offloading) to manage memory efficiency. We configure the per-device batch size to 4 and accumulate gradients over 2 steps, yielding a global batch size of 64. The entire training pipeline is deployed on a single server equipped with 8 $\times$ NVIDIA A100 (80GB) GPUs. The training requires approximately 48 GPU-hours for the 3B model and 72 GPU-hours for the 7B model , averaged per algorithm.

F Diagnostic Annotation Protocol

To ground our analysis of visual reasoning breakdowns in concrete model behavior, we construct a token-level annotated error set using the Qwen2.5-VL-3B model optimized with our standard GRPO baseline. We uniformly sample 200 incorrect rollouts from MathVerse and MathVision, two benchmarks covering diverse forms of visual mathematical reasoning, including diagram understanding, geometric reasoning, symbolic derivation, and chart interpretation. Annotators are asked to identify, for each generated token span, both the source of failure (visual grounding versus textual reasoning) and whether the corresponding operation is executed correctly. This dual-axis evaluation yields a comprehensive four-way taxonomy.

Annotation Principle. Annotations are assigned according to a token's *functional role* within the local reasoning trajectory rather than its isolated lexical meaning. Specifically, annotators determine

Table 4: Key hyperparameters.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	1×10^{-6}
LR Schedule	linear
Epochs	1
Freeze Vision Tower	No
Global Batch Size	64
Per-Device Batch Size	4
Gradient Accumulation	2
GPUs per Node	8
Rollouts per Prompt (G)	4
Rollout Temperature	0.9
Rollout Top-p	1.0
Max Response Length	2048
Reward Signal	accuracy + format
Precision	bfloat16
Gradient Clipping	1.0
DeepSpeed Stage	ZeRO-3

whether a token primarily contributes to image-grounded perception or text-based reasoning under the current decoding context. To improve annotation consistency, labels are first assigned at the semantic-span level and subsequently projected back to individual tokens. Connective tokens are categorized according to the functional role of their subsequent semantic span. For example, the prefix "since angle ADE is $80°$" is labeled as *visually-oriented* because the span performs image-grounded perception, whereas "since Eq. (3) implies" is categorized as *text-oriented* reasoning.

Token Taxonomy. Each generated token within the reasoning chain is classified into exactly one of the following four categories:

- **Correct Visually-oriented Token (P^r).** Tokens that accurately extract or describe image-grounded information. Examples include correctly reading numerical values from a chart, identifying existing geometric relations, or accurately recognizing objects from the visual input.
- **Incorrect Visually-oriented Token (P^e).** Tokens that attempt to describe visual content but directly contradict the provided image. This includes misread numerical values, hallucinated visual elements, or erroneous spatial descriptions.
- **Correct Text-oriented Token (R^r).** Tokens that perform logically valid inference operations conditioned strictly on the preceding context. This encompasses correct algebraic manipulations, valid theorem applications, or intermediate symbolic deductions without requiring new visual evidence.
- **Incorrect Text-oriented Token (R^e).** Tokens corresponding to logically inconsistent or unsupported reasoning steps. Examples include self-contradictory mathematical conclusions, invalid symbolic derivations, or abrupt claims lacking traceable antecedents in the reasoning history.

Annotation Procedure and Agreement. The labeling process follows a strict double-blind protocol. Each sample is independently evaluated by two annotators possessing graduate-level mathematical training. While annotators are provided with the input image, question, full model response, and reference answer, they are explicitly instructed to label tokens based on the internal consistency and visual faithfulness of the generated trajectory, rather than retrofitting labels based on final-answer correctness. Any labeling discrepancies between the two primary annotators are resolved by a third senior referee.

To quantitatively evaluate annotation reliability, we compute Cohen's κ at the token-span level before the final adjudication. The resulting agreement score ($\kappa = 0.85$) demonstrates substantial inter-annotator consistency. The majority of the initial disagreements (approximately 3% of all annotated tokens) are heavily concentrated around the highly ambiguous transition regions between perception and reasoning spans. Crucially, this rigorous diagnostic protocol serves as the standardized foundation for the comparative mechanistic analysis later presented in Section 6.

G Additional Ablation Studies

Ablation on Reweight Strength (α). Table 5 presents the ablation study on the reweight strength α, which controls the intensity of modality-aware advantage adjustment during RL training. Overall, introducing a moderate reweighting penalty consistently improves visual reasoning performance

Table 5: Ablation Study on Reweight Strength α. Bold formatting indicates the highest performance in each column.

Method	WeMath	MathVision	MathVerse	LogicVista	HallusionBench	MME	MMBench	Avg.
GRPO Baseline ($\alpha = 0$)	23.0	21.4	34.4	39.8	60.8	79.9	54.8	44.9
$\alpha = 0.1$	25.7	23.3	34.5	39.8	**63.4**	80.1	56.5	46.2
DyCo-RL ($\alpha = 0.2$)	**26.9**	22.4	36.0	**41.0**	62.4	80.7	**58.2**	**46.8**
$\alpha = 0.3$	25.0	**24.8**	**36.4**	40.2	59.4	**81.7**	57.8	46.5
$\alpha = 0.4$	23.1	19.5	32.4	37.4	59.9	79.0	54.8	43.7
$\alpha = 0.5$	21.0	17.8	31.5	35.3	58.6	78.2	51.0	41.9

Table 6: Comparison between Advantage Reweighting and Reward Shaping. Bold formatting indicates the highest performance in each column.

Method	WeMath	MathVision	MathVerse	LogicVista	HallusionBench	MME	MMBench	Avg.
GRPO Baseline	23.0	21.4	34.4	38.9	60.8	79.9	54.8	44.7
DyCo-RL (Reward Shaping)	24.4	19.7	34.5	37.6	**62.4**	80.3	**61.3**	45.7
DyCo-RL (Adv. Reweighting)	**26.9**	**22.4**	**36.0**	**41.0**	**62.4**	**80.7**	58.2	**46.8**

over the unweighted baseline ($\alpha = 0$). Specifically, $\alpha = 0.2$ achieves the optimal balance, yielding the strongest or near-strongest results across most datasets. This suggests that moderate role assignment effectively encourages visually-oriented tokens to acquire richer visual evidence while preserving stable reasoning dynamics.

Conversely, when α is too small (e.g., $\alpha = 0.1$), the signal is insufficient to meaningfully reshape cross-modal attention behaviors. At the other extreme, excessively large values ($\alpha \geq 0.4$) precipitously degrade performance. We hypothesize that overly aggressive reweighting disrupts the model's intrinsic coordination, forcing attention patterns into overly constrained sub-optima. The consistent trend across benchmarks confirms that balanced modulation is strictly more beneficial than aggressive intervention.

Advantage Reweighting vs. Reward Shaping. As shown in Table 6, we further contrast our token-level advantage reweighting with a straightforward trajectory-level reward shaping baseline. In the latter, the training signal is modified by directly augmenting the binary accuracy reward (R_{acc}) with the aggregated alignment score ($\bar{s}_t$):

$$R' = R_{\text{acc}} + \alpha \bar{s}_t. \tag{12}$$

While this shaping approach occasionally outperforms the vanilla GRPO baseline, its gains are highly inconsistent and overall inferior to advantage reweighting. We attribute this discrepancy to the optimization dynamics of GRPO. Reward shaping perturbs the trajectory-level reward prior to group normalization, which artificially alters the scale and variance of the resulting advantage estimates. This amplification of noise in the group-relative baseline leads to unstable policy updates. Advantage reweighting, by contrast, injects fine-grained guidance after the baseline computation, successfully avoiding these destabilizing effects.

Scaling the Number of Training Rollouts. The number of rollouts per prompt, R, dictates a fundamental trade-off in group-relative RL: a larger R yields lower-variance advantage estimates but linearly inflates training costs. To identify where this trade-off saturates for DyCo-RL, we evaluate $R \in \{4, 8, 16\}$ while holding all other hyperparameters constant. As detailed in Table 7, performance scales monotonically with R, but exhibits sharply diminishing returns. Moving from $R = 4$ to $R = 8$ provides a noticeable lift across nearly all benchmarks, whereas $R = 16$ yields only marginal further increments. Given that the computational overhead scales linearly, doubling the compute budget from $R = 8$ to $R = 16$ is rarely justifiable for fractional score improvements. Consequently, we adopt $R = 4$ as the default configuration to maintain high efficiency, while noting that practitioners with abundant compute budgets can straightforwardly scale R for maximum performance.

H Analysis of Computational Overhead

As detailed in Table 8, DyCo-RL introduces a moderate computational overhead compared to the standard GRPO baseline, reducing overall training throughput by approximately 27%. This cost primarily stems from the token-level Fisher–Rao distance computation and the subsequent role assignment operations at each generation step. Because these mechanisms require fine-grained extraction

Table 7: Ablation Study on the Number of Training Rollouts (R).

Method	WeMath	MathVision	MathVerse	LogicVista	HallusionBench	MME	MMBench	Avg.
DyCo-RL ($R = 4$, Default)	25.9	22.4	**36.0**	41.0	62.4	80.7	58.3	46.7
DyCo-RL ($R = 8$)	27.1	23.0	35.3	41.4	62.7	82.0	59.6	47.3
DyCo-RL ($R = 16$)	**28.0**	**23.5**	**36.0**	**42.3**	**63.6**	**82.5**	**61.3**	**48.2**

Table 8: Analysis of training throughput and computational overhead.

Method	Throughput (samples/sec)	Δ Throughput
GRPO Baseline	0.447	–
DyCo-RL	**0.327**	**-26.8%**

and alignment of cross-modal attention weights throughout the decoding trajectory, they inherently increase training-time memory bandwidth utilization and computation.

Crucially, this overhead is strictly confined to the optimization phase. At inference time, the advantage reweighting pipeline is completely detached, allowing the optimized model to retain the exact generation speed and memory footprint of the vanilla backbone. Given the substantial performance surges across diverse visual and mathematical reasoning benchmarks, this one-time training cost represents a highly favorable trade-off for mitigating cross-modal coordination breakdowns.

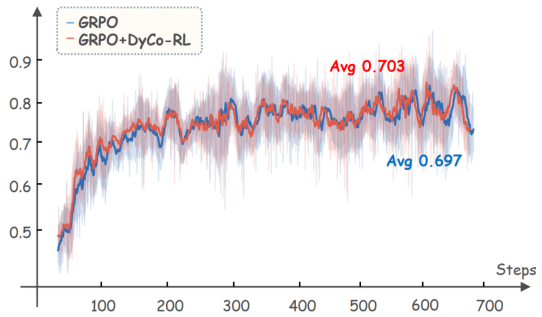


Figure 5: Training dynamics comparing DyCo-RL against the standard GRPO baseline. The curves illustrate the smoothed average accuracy reward across training steps, demonstrating the accelerated convergence and sustained learning signal of our method.

I Analysis of Training Dynamics

To elucidate how DyCo-RL reshapes the optimization landscape, we compare its training trajectory against the standard GRPO baseline with respect to the accuracy reward. For visual clarity, the plotted curves are smoothed using a rolling mean with a window size of 10.

As illustrated in Fig. 5, DyCo-RL demonstrates accelerated convergence during the early stages of training and sustains a consistent performance margin over GRPO. Throughout the entire optimization process, our method maintains a strictly higher average reward. This robust trajectory confirms that role-aware token-level advantage reweighting delivers a more stable and effective learning signal than conventional sequence-level reward broadcasting.

J Extended Generalization Analysis

A critical question regarding token-level advantage reweighting is whether the learned alignment signal overfits to the specific visual-mathematical distribution seen during training. To rigorously assess its out-of-domain robustness, we evaluate DyCo-RL on several additional benchmarks that encompass diverse multimodal reasoning scenarios. These include commonsense visual understanding, real-world image reasoning, chart interpretation, and general multimodal question answering.

As reported in Table 9, DyCo-RL consistently outperforms the standard GRPO baseline across all evaluated zero-shot benchmarks. These results confirm that our proposed dynamic alignment mechanism is not spuriously correlated with the training domain. Rather, by fundamentally addressing the cross-modal coordination bottleneck, the role-aware reweighting strategy generalizes seamlessly across vastly different visual formats, reasoning styles, and task structures.

K Prompt Template

We adopt a GRPO-style reasoning format that requires the model to generate intermediate reasoning within <think> tags and final responses within <answer> tags:

Table 9: Out-of-Domain Generalization Results. Bold formatting indicates the highest performance on each benchmark. DyCo-RL consistently outperforms the baseline across diverse non-mathematical tasks, demonstrating robust cross-domain generalization.

Method	A-OKVQA	RealWorldQA	MMStar	SEED-Bench (IMG)	ChartQA (Test)
GRPO Baseline	82.3	47.0	51.2	68.2	81.0
DyCo-RL	**82.7**	**49.5**	**53.2**	**69.0**	**83.2**

Prompt Template

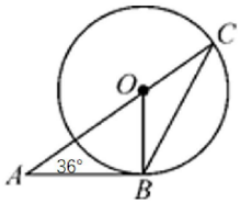
SYSTEM: You are a helpful assistant.

USER:A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively.
Question: {question}

L More Visualization Results

We provide additional qualitative examples in Fig. 6–7, comparing the standard GRPO baseline against our DyCo-RL augmented version on the Qwen2.5-VL-3B architecture. DyCo-RL consistently exhibits role-aligned attention and dynamic cross-modal alternation across diverse visual reasoning cases, further corroborating the findings in the main paper.

As shown in the figure, AB is tangent to circle O at point B, then angle C is equal to ()
Choices: A:36° B:54° C:60° D:27°



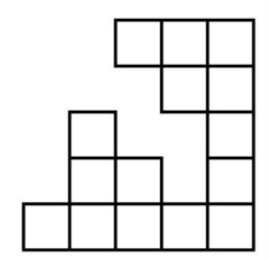
Baseline

<think>\nSince AB is tangent to circle O at point B, angle ABO is a right angle (90°). Given that angle A is 36°, we can find angle OBA by subtracting angle A from 90°, which gives us 54°. Since angle OBA is an exterior angle of triangle OBC, angle C is equal to angle OBA, which is 54°.</think>
<answer>B</answer>

According to the question shown in the image, please directly answer the question and provide the correct option letter, e.g., A, B, C, D.

OURS:

<think>\nSince AB is tangent to circle O at point B, angle ABO is 90 degrees. Given that angle A is 36 degrees, angle OBA can be calculated as 90 - 36 = 54 degrees. Since angle C is an inscribed angle that intercepts the same arc as angle AOB, angle C is half of angle BOA, which is 54 / 2 = 27 degrees.</think>
<answer>D</answer>



Baseline

<think>The big square is made up of 25 small squares. If we count the small squares that are still present in the image, we can determine how many have been lost. There are 16 small squares that are still present.</think>
<answer>9

A big square is made from 25 small squares put together. A few of the small squares have been lost. How many have been lost?

OURS:

<think>The big square is made up of 25 small squares. The shape in the image shows that 15 small squares are visible. Therefore, the number of lost squares is the total number of small squares minus the number of visible squares.
</think>
<answer>10

Figure 6: Visualization Cases on Mathematical and Logical Reasoning.

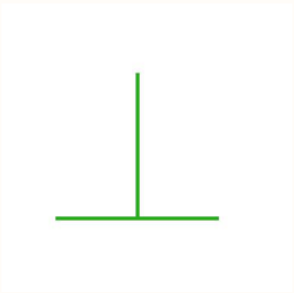
What kind of sport is being played in this image?
A: Soccer
B: Basketball
C: Football (American),
D: Rugby

Baseline

<think> The image shows a group of players running on a field and competing for a ball. The presence of an oval-shaped ball suggests that this is American football rather than another sport.
</think>
<answer>C</answer>

OURS:

<think>The image shows a group of players running on a field, with one player holding a yellow oval-shaped ball. The attire and equipment, including the shorts and the ball, are characteristic of Australian rules football, which is commonly known as \"football\" in Australia and New Zealand.</think>
<answer>D</answer>



Is the vertical line longer than the horizontal line?

Baseline

<think> The image shows two lines: one vertical and one horizontal. The vertical line appears to extend higher than the horizontal line.</think>
<answer> yes </answer>

OURS:

<think>To determine if the vertical line is longer than the horizontal line, I need to compare their lengths visually.The vertical line does not appear longer than the horizontal line.</think>
<answer>No</answer>

Figure 7: Visualization Cases on General Task and Hallucination Reasoning.