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

Anonymous authors
Paper under double-blind review

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 and overlook how that outcome is produced: as a chain-of-thought unfolds, some tokens must extract evidence from the image while others must anchor to the preceding textual context, yet a single sequence-level reward provides no signal about whether each token draws on the modality its role demands. Token-level analyses and controlled interventions reveal that tokens frequently under-attend to their designated modality, and that this coordination breakdown may contribute 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 using 11K examples for RL training demonstrate that the algorithm-agnostic DyCo-RL, when applied to Qwen2.5-VL-3B/7B and Qwen3.5-27B, improves the average performance of all four evaluated RLVR algorithms across seven benchmarks spanning visual-centric and mathematical reasoning. Our code is available at `https://anonymous.4open.science/r/dyco-4122`.

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. Existing RLVR methods, however, provide limited explicit guidance for coordinating visual evidence extraction and textual-context anchoring according to each token's functional role. They either broadcast a single sequence-level advantage uniformly to all tokens (Shao et al., 2024; Zheng et al., 2025), or reinforce visual perception alone (Yang et al., 2025; Wang et al., 2025d; Huang et al., 2025). In both cases, whether a token draws on the modality its current step requires is never supervised.

To uncover the mechanism behind these failures, we examine the rationale generation process at the level of token attention, tracking how each token distributes its attention between visual patches and the textual context (Zhang et al., 2025c; Hassanin et al., 2024; Guo et al., 2022). As illustrated in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

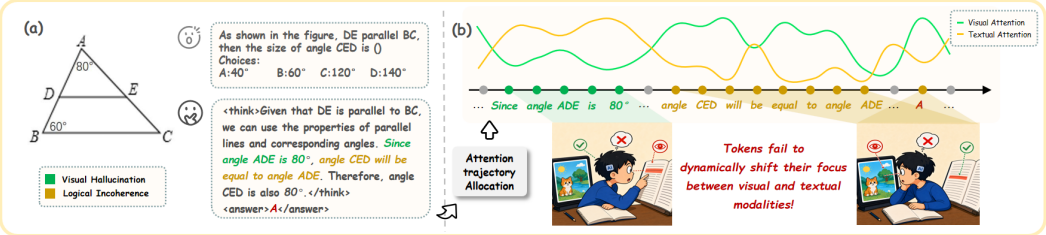


Figure 1: An illustrative example of reasoning failures in Qwen2.5-VL-3B (Bai et al., 2025) optimized via GRPO on ThinkLite-hard-11K (Wang et al., 2025c). (**a**) The model first hallucinates visual features (mistakenly asserting $\angle\text{ADE} = 80°$), and subsequently generates logically incoherent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

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 provide *causal evidence* through controlled intervention experiments. We demonstrate that dynamically amplifying the appropriate modality attention based on a token's functional role can rectify reasoning failures and recover 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 (i.e., visually-oriented, text-oriented, or neutral); a larger distance reflects greater attention redistribution, which we use as a proxy for 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.
- Using only 11K training examples, DyCo-RL improves seven-benchmark average scores across Qwen2.5-VL-3B/7B and Qwen3.5-27B with GRPO, DAPO, SAPO, and GSPO, gaining up to 3.2 points. Additional comparisons show advantages over related methods, while ablations support role-conditioned scoring.

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) derive advantages from sequence-level rewards, applying uniform credit assignment across 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 (Zhang et al., 2025d) incorporates dense step-level rewards to evaluate the accuracy

and logical consistency of intermediate reasoning steps. Finally, at the token level, fine-grained optimization has been explored for preference alignment in LLMs (Zeng et al., 2024; Lin et al., 2025) and entropy-based token selection in RLVR (Wang et al., 2025b). 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 clustering. Despite these advances, explicit guidance for credit assignment that jointly accounts for visual evidence extraction and reasoning over the textual context remains limited. Our work addresses this gap by reweighting token-level advantages according to role-specific modality alignment.

Cross-Modal Attention Analysis in MLLMs. Attention patterns serve as a crucial window into the cross-modal interactions of MLLMs. One line of research explores inference-time interventions, including attention amplification, attention-guided visual cropping, and visual-aware decoding (Liu et al., 2024a; Zhang et al., 2025c; Wang et al., 2025a), to improve visual perception and mitigate hallucinations. Another line of research investigates attention from an interpretability perspective. Recent studies show that models can attend to relevant image regions even when producing incorrect answers (Zhang et al., 2025c), and that visual attention may decline during extended reasoning (Sun et al., 2025), motivating closer examination of attention allocation throughout generation. Attention has also been incorporated into RL through reward design and token selection (Jian et al., 2025; Jiao et al., 2026). Building on these insights, DyCo-RL uses within-modality attention shifts as a proxy for functional role assignment, translating attention dynamics into role-conditioned guidance for RLVR optimization.

3 PRELIMINARIES

3.1 MODALITY-SPECIFIC ATTENTION DECOMPOSITION

In autoregressive MLLMs, the decoder attends to both visual and textual tokens at each generation step. Let $\mathcal{I}_{\text{vis}}$ denote the indices of visual tokens and $\mathcal{I}_{\text{txt}}$ the indices of non-system textual tokens, including the user prompt, associated special tokens, and all generated tokens. At the last decoder layer, the attention weight at decoding step t to a context position $j \le t$, averaged over heads, is

$$a_{t,j} = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \frac{\exp(e^h_{tj})}{\sum_{k \le t} \exp(e^h_{tk})}, \qquad e^h_{tj} = \frac{(\mathbf{q}^h_t)^\top \mathbf{k}^h_j}{\sqrt{d}}, \tag{1}$$

where $\mathcal{H}$ denotes the set of attention heads, $\mathbf{q}^h_t$ and $\mathbf{k}^h_j$ are the query and key vectors of head h, respectively, and d is the head dimension. The *modality-specific attention mass* of step t on modality $m \in \{\text{vis}, \text{txt}\}$ is

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

which measures how much attention the generated token allocates to modality m in total.

3.2 GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Given an input prompt x, the current policy $\pi_{\theta_{\text{old}}}$ (i.e., the policy before the update) samples 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 without 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 by maximizing a clipped surrogate objective, in which the scalar advantage $\hat{A}_i$ is applied to every token position of the response:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{X},\ \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \Big(\rho_{i,t}\, \hat{A}_i,\ \text{clip}\big(\rho_{i,t}, 1-\epsilon, 1+\epsilon\big) \hat{A}_i \Big) \right], \tag{4}$$

where $\rho_{i,t} = \pi_\theta(o_{i,t} \mid x, o_{i,<t}) \,/\, \pi_{\theta_{\text{old}}}(o_{i,t} \mid x, o_{i,<t})$ is the importance sampling ratio of the t-th token, and the KL regularization term is omitted for brevity.

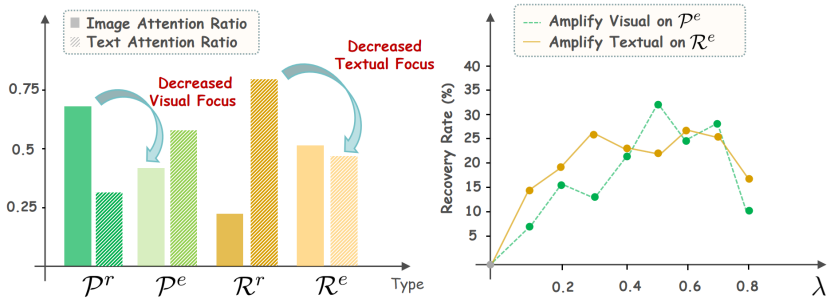


Figure 2: **(Left)** Correlation analysis of modality-specific attention. $\mathcal{P}/\mathcal{R}$ denote visually/text-oriented tokens, and superscripts r/e indicate correct/erroneous states. Attention ratios are computed as $r_t^m/(r_t^{\text{vis}} + r_t^{\text{txt}})$. **(Right)** Causal intervention 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.

Since $\hat{A}_i$ is shared by all tokens of o_i, a token that reads a value from the image and a token that derives a conclusion from the preceding text receive the same learning signal, regardless of whether each attends to the modality its step requires.

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 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) 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 ratio $r_t^m/(r_t^{\text{vis}} + r_t^{\text{txt}})$ 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. Additional correlation analysis is provided in Appendix G.

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, at each intervened decoding step t, we enhance attention toward the target modality independently within each head of the last decoder layer:

$$\tilde{a}_{t,j}^h = \frac{\exp(e_{tj}^h)\big(1 + \lambda \mathbf{1}_{j \in \mathcal{M}}\big)}{\sum_{k \le t} \exp(e_{tk}^h)\big(1 + \lambda \mathbf{1}_{k \in \mathcal{M}}\big)}. \tag{5}$$

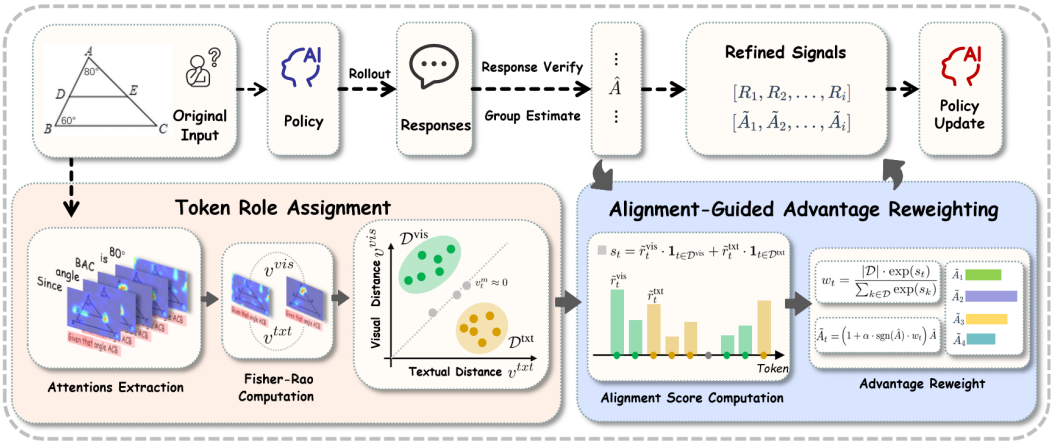


Figure 3: **Overall Pipeline of DyCo-RL.** Instead of broadcasting the sequence-level advantage $\hat{A}_i$ to all tokens, DyCo-RL assigns each token a functional role via the Fisher–Rao distance between consecutive within-modality attention distributions, scores how well the token's attention matches that role, and uses the score to reweight $\hat{A}_i$ at the token level, yielding $\tilde{A}_{i,t}$ for the policy update.

Here, e^h_{tj} is the attention logit defined in Eq. (1), $\tilde{a}^h_{t,j}$ is the modified attention weight of head h, and $\lambda > 0$ controls the enhancement strength. The modified weights are normalized separately within each head. Details of the replay protocol and control conditions are provided in Appendix H.

As shown in Fig. 2, moderate enhancement can recover correct answers on previously erroneous samples for both functional roles, with a higher peak recovery rate for visually-oriented interventions. Excessively large λ, however, degrades performance, indicating that over-amplification disrupts the original attention balance. The role-aware intervention also outperforms the controls at most tested strengths (Appendix H). Together with the correlation analysis, these findings provide causal evidence that role-aligned attention modulation can help correct 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

Acquiring new information from a modality may involve redistributing attention across positions within that modality. We therefore characterize a token's functional role through changes in its within-modality attention distribution between consecutive steps. Total attention mass measures how much attention a modality receives, but does not distinguish persistent attention patterns, such as attention sinks (Queipo-de Llano et al., 2026), from step-specific shifts in focus. We use within-modality redistribution to infer functional roles and, once a role is assigned, use modality-level attention mass to score the attention allocated to the corresponding modality.

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:

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

For the textual modality, we restrict both consecutive distributions to their shared prefix and independently renormalize them before computing the distance. A high v^m_t indicates substantial within-modality attention redistribution, which we use as a proxy for active information acquisition, whereas $v^m_t \approx 0$ indicates little change in the attention distribution. Under the square-root embedding $p \mapsto 2\sqrt{p}$, the Fisher–Rao distance measures spherical geodesic separation between attention distributions, providing a symmetric measure bounded in $[0, \pi]$.

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

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

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 proposed token role assignment, we translate modality alignment into fine-grained RL optimization guidance. To derive token-level operations, we temporarily drop the response index i and denote the response length by T. For each modality m, we min–max normalize its attention mass across the response: $\tilde{r}_t^m = \frac{r_t^m - \min_k r_k^m}{\max_k r_k^m - \min_k r_k^m}$, where $t, k \in \{1, \ldots, T\}$. The alignment score selects the normalized attention mass corresponding to the token's assigned functional role:

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

Let $\mathcal{D} = \mathcal{D}^{\text{vis}} \cup \mathcal{D}^{\text{txt}}$ denote the set of non-neutral tokens. Tokens in $\mathcal{D}$ have their advantage reweighted by the alignment score, while neutral tokens retain the sequence-level advantage:

$$\tilde{A}_t = \begin{cases} \Big(1 + \alpha \cdot \operatorname{sgn}(\hat{A}) \cdot w_t\Big)\, \hat{A}, & t \in \mathcal{D}, \\ \hat{A}, & t \notin \mathcal{D}, \end{cases} \qquad w_t = \frac{|\mathcal{D}| \cdot \exp(s_t)}{\sum_{k \in \mathcal{D}} \exp(s_k)}, \quad t \in \mathcal{D}. \tag{10}$$

where $\alpha > 0$ controls the reweighting strength and $|\mathcal{D}|$ normalizes w_t to unit mean over $\mathcal{D}$; a centered variant removing the resulting common shift is evaluated in Table 17.

The reweighting thus reinforces role-aligned behavior by amplifying positive advantages and attenuating negative ones more strongly for tokens with higher alignment scores, encouraging the model to read the image or use the preceding context when the role requires it. Substituting $\tilde{A}_{i,t}$ for $\hat{A}_i$ in Eq. (4), the DyCo-RL objective with GRPO as the base algorithm is

$$\mathcal{J}_{\text{DyCo-RL}}(\theta) = \mathbb{E}_{x \sim \mathcal{X},\, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(\rho_{i,t}\, \tilde{A}_{i,t},\ \text{clip}\big(\rho_{i,t}, 1-\epsilon, 1+\epsilon\big)\, \tilde{A}_{i,t} \Big) \right]. \tag{11}$$

5 Experiments

5.1 Experimental Settings

Models, Data, and Baselines. We integrate DyCo-RL into Qwen2.5-VL-3B/7B and Qwen3.5-27B, utilizing the ThinkLite-hard-11K dataset for training. Qwen2.5-VL is a commonly used open-source backbone in recent MLLM reinforcement-learning studies, enabling comparisons under a stable architecture and helping isolate the effect of the optimization method. We further include Qwen3.5-27B (Qwen Team, 2026) as a newer and substantially larger backbone to examine whether the benefits of DyCo-RL generalize across model generations and scales, rather than being specific to Qwen2.5-VL. ThinkLite-hard-11K provides a compact yet challenging set of difficulty-curated multimodal reasoning samples selected via MCTS from a substantially larger data pool, making it well suited for resource-efficient RLVR 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. Due to limited computational resources, we train on the 11K samples for one epoch, with the main results reported using random seed 42. Multi-seed results are provided in Appendix K. Models are trained with a learning rate of 1e-6, 4 rollouts per prompt, 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 $\tau = 0.05$. Algorithm-specific hyperparameters of all RLVR baselines follow their official implementations. 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 (Xiao et al., 2024), HallusionBench (Guan et al., 2024), MME (Fu et al., 2023), and MMBench (Liu et al., 2024b). Comprehensive evaluation configurations are detailed in Appendix D.

Table 1: Main Results across four RLVR algorithms and three 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	MMBench	LogicVista	HallusionBench	MME	Avg.
Qwen2.5-VL-3B								
Zero-Shot	20.0	23.0	33.8	71.4	34.5	61.3	76.9	45.8
GRPO	23.0	21.4	34.4	72.8	39.8	60.8	79.9	47.4
GRPO + DyCo-RL	**25.9**	**22.4**	**36.0**	**76.2**	**41.2**	**62.4**	**80.7**	**49.3**
DAPO	**27.4**	17.1	34.0	73.8	39.8	60.7	81.0	47.7
DAPO + DyCo-RL	26.9	**23.0**	**36.6**	**75.6**	**41.4**	**62.7**	**82.1**	**49.8**
SAPO	31.5	**20.1**	35.0	75.4	39.4	59.7	**82.1**	49.0
SAPO + DyCo-RL	**32.0**	19.7	**35.3**	**77.2**	**44.5**	**63.9**	81.1	**50.5**
GSPO	**27.2**	19.7	34.0	76.6	35.8	**62.4**	79.6	47.9
GSPO + DyCo-RL	26.0	**24.0**	**34.1**	**84.0**	**37.4**	62.0	**81.9**	**49.9**
Qwen2.5-VL-7B								
Zero-Shot	33.2	23.7	43.8	82.9	44.5	65.0	82.4	53.6
GRPO	38.7	26.6	47.4	78.0	46.1	69.2	83.3	55.6
GRPO + DyCo-RL	**40.5**	**29.3**	**49.5**	**83.8**	**48.5**	**71.6**	**84.9**	**58.3**
DAPO	36.4	23.0	44.5	79.2	42.3	70.0	**86.3**	54.5
DAPO + DyCo-RL	**37.5**	**24.3**	**45.2**	**85.8**	**43.8**	**71.0**	86.0	**56.2**
SAPO	38.5	**26.6**	47.0	82.4	42.3	**71.0**	85.6	56.2
SAPO + DyCo-RL	**40.7**	24.7	**47.8**	**84.3**	**47.7**	69.2	**86.0**	**57.2**
GSPO	39.4	26.6	42.4	75.8	46.5	71.4	81.6	54.8
GSPO + DyCo-RL	**39.9**	**27.6**	**46.5**	**85.5**	**48.3**	**73.2**	**84.8**	**58.0**
Qwen3.5-27B								
Zero-Shot	55.0	33.6	57.7	84.5	43.6	71.4	87.9	62.0
GRPO	**61.4**	39.8	57.9	82.6	45.2	73.1	88.5	64.1
GRPO + DyCo-RL	61.2	**41.1**	**61.0**	**86.4**	**49.0**	**74.3**	**89.1**	**66.0**
DAPO	**60.5**	38.5	59.0	83.8	45.4	71.0	87.1	63.6
DAPO + DyCo-RL	59.3	**40.1**	**59.9**	**85.6**	**48.8**	**72.9**	**89.4**	**65.1**
SAPO	58.7	36.2	58.4	85.0	45.4	71.3	**88.4**	63.3
SAPO + DyCo-RL	**61.0**	**38.5**	**58.8**	**87.1**	**46.1**	**71.8**	**88.4**	**64.5**
GSPO	60.0	37.2	60.0	85.1	**47.7**	72.1	88.6	64.4
GSPO + DyCo-RL	**61.9**	**39.5**	**60.2**	**87.3**	47.0	**75.2**	**88.9**	**65.7**

5.2 Main Results

Table 1 presents evaluation results on seven benchmarks across three backbone configurations (Qwen2.5-VL-3B/7B and Qwen3.5-27B) and four RLVR algorithms. DyCo-RL improves the average performance of every evaluated baseline across all three configurations, supporting its applicability across model scales and optimization algorithms.

Algorithm-Agnostic Generality. DyCo-RL yields average performance gains across GRPO, DAPO, SAPO, and GSPO without requiring algorithm-specific modifications. Despite differences in their optimization objectives, the improvements hold under every evaluated backbone configuration. This transferability across clip-based, sequence-level, and soft-adaptive optimization paradigms suggests that alignment-guided advantage reweighting is compatible with diverse RLVR objectives and can serve as a broadly applicable enhancement within the evaluated pipelines.

Balanced Improvements Across Reasoning and Perception. The gains span both mathematical and visual-centric reasoning benchmarks. Improvements on visual tasks such as HallusionBench and MME are accompanied by gains on MathVision and MathVerse in multiple configurations. Although individual benchmarks exhibit occasional regressions, the overall results support the value of jointly considering visual and textual demands when optimizing multimodal reasoning.

Scalability Across Model Sizes and Generations. Beyond Qwen2.5-VL-3B, DyCo-RL improves average performance on Qwen2.5-VL-7B and Qwen3.5-27B across all four algorithms. These improvements extend across model sizes and backbone generations, supporting the applicability of role-aware advantage reweighting beyond a single backbone configuration.

5.3 Comparison with Visual Reasoning Methods

Beyond general RLVR algorithms, we compare DyCo-RL with Attention Intervention and two representative methods with publicly available implementations, PAPO-G and VPPO, using Qwen2.5-VL-3B as the common backbone. Attention Intervention applies the head-wise enhancement in Eq. (5) with $\lambda = 0.2$ to the zero-shot model at inference time, without RL fine-tuning. Target modalities are determined by the Fisher–Rao roles in Eq. (8). PAPO-G (Wang et al., 2025d) augments GRPO with a

Table 2: Comparison with visual reasoning methods on Qwen2.5-VL-3B.

Method (Qwen2.5-VL-3B)	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
Zero-Shot	20.0	23.0	33.8	71.4	34.5	61.3	76.9	45.8
GRPO	23.0	21.4	34.4	72.8	39.8	60.8	79.9	47.4
Attention Intervention	24.4	21.7	34.5	74.0	40.3	**62.7**	79.0	48.1
PAPO-G	22.8	21.4	34.6	74.3	40.3	61.8	80.3	47.9
VPPO	**26.9**	22.0	35.8	75.1	39.6	61.2	78.5	48.4
GRPO + DyCo-RL	25.9	**22.4**	**36.0**	**76.2**	**41.2**	62.4	**80.7**	**49.3**

Table 3: Effectiveness of Fisher–Rao Role Assignment compared to alternative assignment strategies.

Method	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
GRPO (Baseline)	23.0	21.4	34.4	72.8	39.8	60.8	79.9	47.4
+ Random	23.5	20.1	33.1	72.4	36.7	59.7	79.3	46.4
+ Reverse	24.2	19.4	34.0	70.6	36.0	58.6	78.0	45.8
+ Entropy	25.0	21.1	34.4	74.5	40.3	61.2	**81.1**	48.2
+ KL Divergence	24.7	21.4	35.4	75.3	40.7	**62.5**	80.4	48.6
+ DyCo-RL (Ours)	**25.9**	**22.4**	**36.0**	**76.2**	**41.2**	62.4	80.7	**49.3**

KL-based implicit perception loss. VPPO (Huang et al., 2025) reweights trajectory-level advantages by visual dependency and restricts policy updates to perceptually pivotal tokens. These comparisons evaluate role-aware advantage reweighting against inference-time intervention, perception-focused objectives, and visually guided token-level optimization.

As shown in Table 2, DyCo-RL achieves the highest average score (49.3), exceeding PAPO-G (47.9), VPPO (48.4), and Attention Intervention (48.1). It outperforms PAPO-G on all seven benchmarks, including gains of 3.1 points on WeMath and 1.9 points on MMBench. It surpasses VPPO on six benchmarks, with gains of 1.1 points on MMBench, 2.2 points on MME, and 1.6 points on LogicVista. These results support incorporating dynamic visual–textual coordination into training, with benefits over the evaluated inference-time and perception-focused methods.

5.4 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. Ablation experiments are conducted on the Qwen2.5-VL-3B model using the GRPO baseline. Additional ablation results are provided in Appendix J.

Effectiveness of Fisher–Rao Role Assignment. We compare DyCo-RL against four alternative token role assignment strategies (Table 3). *Random* assigns roles arbitrarily and lowers the average score by 1.0 points relative to GRPO, while *Reverse* inverts our assignments and yields the lowest average score. These controls support the importance of informed role assignment and indicate that reversing the assignment direction reduces the effectiveness of reweighting.

We further compare two heuristic baselines: *Entropy*, which measures the uncertainty of the within-modality attention distribution at a single step, and *KL*, which captures changes between consecutive distributions. Entropy provides a static snapshot, whereas KL incorporates temporal changes but is asymmetric. Fisher–Rao instead provides a symmetric, geometrically principled distance on the probability simplex, characterizing how attention is redistributed within each modality between consecutive decoding steps. DyCo-RL achieves the highest average performance among the evaluated variants, supporting the use of Fisher–Rao-based role assignment for token-level advantage reweighting.

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 based on a token's assigned role ($\mathcal{D}^{\text{vis}}$ or $\mathcal{D}^{\text{txt}}$), we use a fixed modality's min–max normalized attention mass for all tokens in $\mathcal{D}$: *(i) + visual attention score* (setting $s_t = \tilde{r}_t^{\text{vis}}$), and *(ii) + text attention score* (setting $s_t = \tilde{r}_t^{\text{txt}}$). As shown in Table 4, optimizing either modality alone can improve certain benchmarks but degrades others, and neither consistently surpasses GRPO. This indicates that one-sided modality optimization is insufficient. In contrast, DyCo-RL jointly accounts for both modalities and selectively reweights tokens according to their functional roles, yielding the highest overall performance.

Table 4: Effectiveness of Alignment-Guided Reweighting.

Method	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
GRPO (Baseline)	23.0	21.4	34.4	72.8	39.8	60.8	79.9	47.4
+ Visual Attn Score ($\tilde{r}_t^{\text{vis}}$)	24.5	**23.0**	32.1	74.4	37.4	62.0	78.1	47.4
+ Text Attn Score ($\tilde{r}_t^{\text{txt}}$)	**26.0**	22.4	35.4	70.7	37.8	59.4	78.4	47.2
+ DyCo-RL (Ours)	25.9	22.4	**36.0**	**76.2**	**41.2**	**62.4**	**80.7**	**49.3**

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, with annotators blinded to model identity. Based on this curated evaluation set, we investigate how our alignment-guided reweighting fundamentally reshapes cross-modal reasoning dynamics.

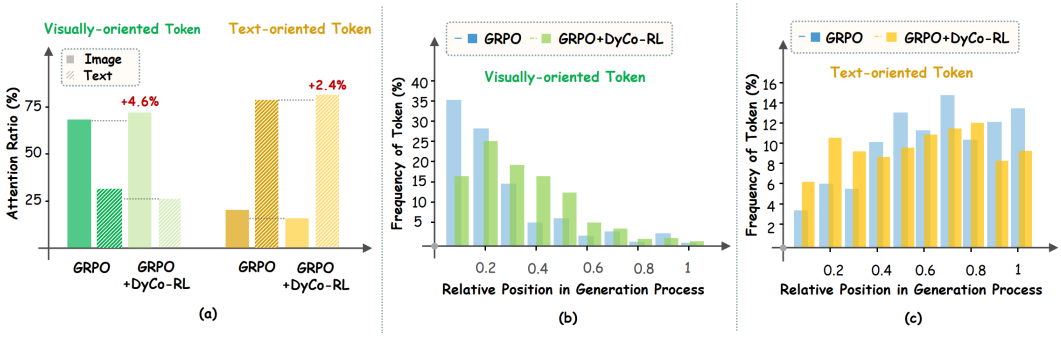


Figure 4: Effectiveness of DyCo-RL in dynamic cross-modal coordination. (a) Alignment between assigned token roles and modality attention ratios; (b) Distribution of visually-oriented tokens across the generation process; (c) Distribution of text-oriented tokens across the generation process.

Strengthening Role-Aligned Attention. We first evaluate the attention allocations across the annotated tokens. Attention ratios are computed as in Fig. 2. As illustrated in Fig. 4(a), DyCo-RL strengthens the alignment between a token's functional role and its actual modality attention. Specifically, compared with GRPO, DyCo-RL exhibits a higher image attention ratio for visually-oriented tokens and a higher text attention ratio for text-oriented tokens. The shift is partly expected by construction, since the objective rewards role-aligned attention. This trend confirms that concentrating the learning signal on role-aligned tokens shifts attention toward the designated modality.

Reshaping Temporal Dynamics. Fig. 4(b–c) shows the temporal distribution of each token role, measured as the fraction of its tokens in each relative-position bin. Under GRPO, visually-oriented tokens are concentrated at the beginning, with a small proportion appearing after the midpoint, while text-oriented tokens become increasingly frequent toward the end. DyCo-RL reduces this separation: visually-oriented tokens have a greater presence in the middle stages, while text-oriented tokens appear earlier and are distributed more evenly. This more interleaved distribution suggests a reduced tendency to "perceive first, reason later" and, together with the stronger role–attention alignment in Fig. 4(a), supports dynamic cross-modal coordination as the CoT unfolds.

7 Conclusion

We identify cross-modal coordination breakdowns as a contributing factor to visual reasoning errors in RLVR, where visually- and text-oriented tokens can 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 improves the aggregate performance of four representative baselines across seven diverse benchmarks, supporting explicit token-level coordination as an effective approach to improving multimodal reasoning.

AI USAGE STATEMENT

An LLM assisted with manuscript and code revision, and literature checking. The authors retain full responsibility for the research decisions, experimental results, and final manuscript.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

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

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

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025. URL https://arxiv.org/abs/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. 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)*, pp. 14375–14385, June 2024.

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. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3):331–368, 2022. doi: 10.1007/s41095-022-0271-y.

Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad Shahbaz Khan, and Ajmal Mian. Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, 108:102417, 2024. ISSN 1566-2535. doi: 10.1016/j.inffus.2024.102417. URL https://www.sciencedirect.com/science/article/pii/S1566253524001957.

Feijuan He, Yaxian Wang, Xianglin Miao, and Xia Sun. Interpretable visual reasoning: A survey. *Image and Vision Computing*, 112:104194, 2021. ISSN 0262-8856. doi: 10.1016/j.imavis.2021.104194. URL https://www.sciencedirect.com/science/article/pii/S0262885621000998.

Wei He, Zhiheng Xi, Wanxu Zhao, Xiaoran Fan, Yiwen Ding, Zifei Shan, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Distill visual chart reasoning ability from LLMs to MLLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

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

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9251–9270. Association for Computational Linguistics, November 2025. doi: 10.18653/v1/2025.emnlp-main.470. URL https://aclanthology.org/2025.emnlp-main.470/.

Zhengbo Jiao, Shaobo Wang, Zifan Zhang, Wei Wang, Bing Zhao, Hu Wei, and Linfeng Zhang. Credit where it is due: Cross-modality connectivity drives precise reinforcement learning for MLLM reasoning, 2026. URL https://arxiv.org/abs/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. Explain before you answer: A survey on compositional visual reasoning, 2025. URL https://arxiv.org/abs/2508.17298.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37906–37918. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lin25j.html.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *Computer Vision – ECCV 2024*, 2024a. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/10933.pdf.

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

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, Yifan Zhang, et al. 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)*, pp. 20023–20070, 2025.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Ahmad Rezaei, Mohsen Gholami, Saeed Ranjbar Alvar, Kevin Cannons, Mohammad Asiful Hossain, Zhou Weimin, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. CPPO: Contrastive perception for vision language policy optimization. *arXiv preprint arXiv:2601.00501*, 2026.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pp. 146–162, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20073-1. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/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. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Yan Shu, Chi Liu, Robin Chen, Derek Li, and Bryan Dai. Fleming-VL: Towards universal medical visual reasoning with multimodal LLMs, 2025. URL https://arxiv.org/abs/2511.00916.

Yan Shu, Bin Ren, Zhitong Xiong, Xiao Xiang Zhu, Begüm Demir, Nicu Sebe, and Paolo Rota. TerraScope: Pixel-grounded visual reasoning for Earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16712–16722, 2026.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long CoT reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5158–5171. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.acl-long.257. URL https://aclanthology.org/2025.acl-long.257/.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? Dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/24079b91da7257cb78805262996152b8-Paper-Conference.pdf.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2506.01939.

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

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

xAI. RealWorldQA: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26.

Yijia Xiao, Edward Sun, Di Liu, and Wei Wang. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts, 2024. URL https://arxiv.org/abs/2407.04973.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-Back: Implicit visual re-focusing in MLLM reasoning, 2025. URL https://arxiv.org/pdf/2505.07889.

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, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen.

MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of CVPR*, 2024.

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

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

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025b.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://arxiv.org/abs/2502.17422.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, October 2025d. URL https://arxiv.org/abs/2503.12937.

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

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. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

A Overview of Appendix

- B: **Limitations**.
- C: **Broader Impact**.
- D: **Details of Experimental Setup.**
- E: **Implementation Details**.
- F: **Diagnostic Annotation Protocol**.
- G: **Additional Correlation Analysis**.
- H: **Details of Causal Intervention Protocol and Additional Controls**.
- I: **Additional Analysis of Token Role Assignment**.
- J: **Additional Ablation Studies**.
- K: **Multi-Seed Evaluation**.
- L: **Analysis of Computational Overhead**.
- M: **Analysis of Training Dynamics**.
- N: **Extended Generalization Analysis**.
- O: **Prompt Template**.
- P: **More Visualization Results**.

B Limitations

Our work has several limitations. First, although DyCo-RL introduces no additional trainable parameters, extracting per-token attention statistics and computing Fisher–Rao distances during rollout incur additional training time and GPU memory overhead. To limit memory usage, we extract attention statistics only from the last decoder layer, leaving the benefits of incorporating signals from other layers unexplored. Second, due to limited computational resources, we train for a single epoch on ThinkLite-hard-11K and evaluate only Qwen-family backbones. Generalization to other training datasets and model families remains to be verified.

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 (He et al., 2025), long-video reasoning (Zhang et al., 2025b), remote sensing reasoning (Shu et al., 2026), and medical image analysis (Shu et al., 2025). 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 may benefit multimodal reasoning applications through improved cross-modal coordination. However, role-aligned attention does not guarantee faithful or correct reasoning, and models may still produce misleading outputs. Application-specific validation remains necessary before deployment. 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) uses a length-normalized sequence-level importance ratio and sequence-level clipping. To incorporate DyCo-RL's token-specific advantages, we adopt its GSPO-token variant, whose stop-gradient construction preserves the sequence-level ratio while routing gradients through the corresponding token likelihoods. The reweighted advantages are applied to the token-wise surrogate terms before clipping and subsequent averaging.

Evaluation Benchmarks. We evaluate DyCo-RL on seven benchmarks covering mathematical reasoning and visual-centric reasoning to assess visual grounding and reasoning consistency. We follow the benchmark-specific evaluation procedures implemented in VLMEvalKit (Duan et al., 2024), using the splits and metrics specified below.

Mathematical Reasoning.

- **WeMath** (Qiao et al., 2025): A hierarchical visual mathematics benchmark that decomposes composite problems into subproblems associated with individual knowledge concepts, enabling diagnosis of knowledge gaps and reasoning failures. We evaluate on its `testmini` split, containing $1{,}740$ questions, and report the Strict score computed by VLMEvalKit.
- **MathVision** (Wang et al., 2024): A visual mathematical reasoning benchmark comprising $3{,}040$ competition problems across 16 mathematical disciplines. Our evaluation uses the 304-question `testmini` subset, registered as `MathVision_MINI` in VLMEvalKit, and reports final-answer accuracy.
- **MathVerse** (Zhang et al., 2024): A diagram-based mathematical reasoning benchmark with controlled variations in the information provided through text and images. We evaluate on `MathVerse_MINI`, which contains 788 problems presented in five versions: Text Dominant, Text Lite, Vision Intensive, Vision Dominant, and Vision Only, yielding $3{,}940$ evaluation instances. We compute final-answer accuracy for each version.

Visual-centric Reasoning.

- **MMBench** (Liu et al., 2024b): A multiple-choice multimodal benchmark covering diverse perception and reasoning abilities. We use the English development split of v1.0 (`MMBench_DEV_EN`) and report circular accuracy, which evaluates consistency across permutations of the answer options.
- **LogicVista** (Xiao et al., 2024): A visual logical reasoning benchmark covering inductive, deductive, numerical, spatial, and mechanical reasoning. We evaluate the 447 questions provided by VLMEvalKit under `LogicVista` and report overall answer accuracy.
- **HallusionBench** (Guan et al., 2024): A diagnostic benchmark designed to probe visual illusions and language-prior-induced hallucinations through paired image–question controls. We evaluate the 951 instances provided by VLMEvalKit and report answer-level accuracy (aAcc), the proportion of individual responses that correctly answer the corresponding yes/no questions.

- **MME** (Fu et al., 2023): A comprehensive evaluation suite covering 14 perception and cognition tasks, including object recognition, OCR, and commonsense reasoning. We use the standard `MME` evaluation set and its official scoring procedure: each task receives the sum of answer accuracy (Acc) and paired-question accuracy (Acc+), with task scores aggregated within the perception and cognition categories.

Evaluation Protocol. We perform evaluation using the VLM-R1 interface in VLMEvalKit (Duan et al., 2024). Prompts request reasoning and final answers within `<think>` and `<answer>` tags, respectively. The answer-tag content is extracted when available; otherwise, the full response is retained. Each baseline and its DyCo-RL counterpart share the same prompting, answer-processing, and decoding settings, with a decoding temperature of 0.1.

For WeMath, MathVision, MathVerse, and LogicVista, the configured answer-extraction model is `gpt-4o-mini-2024-07-18`, which also performs answer judging where required by the benchmark-specific evaluator. For MMBench, HallusionBench, and MME, VLMEvalKit uses rule-based answer extraction with `gpt-3.5-turbo-0125` configured as the auxiliary extractor when rule-based extraction is insufficient. For MME, we normalize the combined perception and cognition score S_{MME} to a 0–100 scale as $100 \times S_{\mathrm{MME}}/2800$. The overall average is the arithmetic mean of the seven benchmark scores.

We additionally compare answer-extraction failure rates in Table 5, counting answers that remain unresolved after benchmark-specific answer processing. For each model scale, we average these rates equally across the four RLVR algorithms and a common set of benchmarks, comparing each algorithm with and without DyCo-RL. The comparable average extraction failure rates suggest that the observed performance gains are unlikely to be primarily driven by differences in answer extractability.

Table 5: Average answer-extraction failure rates (%) across four RLVR algorithms and seven benchmarks.

Model	Algorithms (Avg.)	Algorithms + DyCo-RL (Avg.)
Qwen2.5-VL-3B	2.02	2.11
Qwen2.5-VL-7B	1.67	1.31

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 6, all models are trained for a single epoch on the ThinkLite-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 follow the settings of the corresponding official implementations unless explicitly stated otherwise.

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. Rollouts are generated using Hugging Face Transformers' `generate`. Baseline training for one epoch requires approximately 54 GPU-hours for Qwen2.5-VL-3B, 72 GPU-hours for Qwen2.5-VL-7B, and 200 GPU-hours for Qwen3.5-27B, averaged across the evaluated baseline algorithms. Additional analysis of computational overhead is reported in Appendix L.

Table 6: Key hyperparameters.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	1×10^{-6}
LR Schedule	linear
Epochs	1
Freeze Vision Tower	No
Global Batch Size (rollout sequences)	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

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 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. Only tokens within spans expressing visual grounding or reasoning operations are annotated, and each annotated token is assigned to exactly one of the following four categories:

- **Correct Visually-oriented Token ($\mathcal{P}^r$).** 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 ($\mathcal{P}^e$).** 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 ($\mathcal{R}^r$).** 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 ($\mathcal{R}^e$).** 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. Each sample is independently evaluated by two external 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. These manual labels are used only for diagnostic analysis and are not involved in DyCo-RL training.

To quantitatively evaluate annotation reliability, we compute Cohen's κ at the token-span level before the final adjudication. The resulting agreement score ($\kappa = 0.85$) indicates high inter-annotator agreement. The majority of the initial disagreements (approximately 3% of all annotated tokens) are concentrated around ambiguous transitions between perception and reasoning spans.

Annotation Statistics. Table 7 reports the token counts for the four annotation groups. The correct-token groups, $\mathcal{P}^r$ and $\mathcal{R}^r$, contain tokens from locally correct spans within the same 200 erroneous rollouts.

Table 7: Token counts in the 200 annotated erroneous rollouts.

Group	Functional role	Correctness	Tokens
$\mathcal{P}^r$	Visually-oriented	Correct	1,533
$\mathcal{P}^e$	Visually-oriented	Erroneous	457
$\mathcal{R}^r$	Text-oriented	Correct	3,624
$\mathcal{R}^e$	Text-oriented	Erroneous	1,261
Total			6,875

G Additional Correlation Analysis

To examine whether the attention gaps in Fig. 2 can be explained by differences in generation position, we compare correct and erroneous tokens at matched generation positions, separately for visually-oriented and text-oriented roles. We use the modality-specific attention ratio $r_t^m/(r_t^{\mathrm{vis}} + r_t^{\mathrm{txt}})$, with $m = \mathrm{vis}$ for visually-oriented tokens and $m = \mathrm{txt}$ for text-oriented tokens. For each role, we report the difference between correct and erroneous tokens. Positive values indicate greater attention to the role-relevant modality for correct tokens.

Table 8: Attention ratio differences between correct and erroneous tokens at matched generation positions.

Functional role	Target modality	Difference
Visually-oriented ($\mathcal{P}$)	Image	$+0.13$
Text-oriented ($\mathcal{R}$)	Textual context	$+0.12$

As shown in Table 8, the differences remain positive for both roles after position matching: correct visually-oriented tokens allocate more attention to the image, while correct text-oriented tokens allocate more attention to the textual context. These descriptive results preserve the direction of the attention–correctness association observed in Fig. 2.

H Details of Causal Intervention Protocol and Additional Controls

We conduct diagnostic attention interventions on the erroneous responses annotated in Appendix F. These experiments examine whether directing attention toward the modality associated with an erroneous span can recover the final answer.

Span-Level Replay and Stopping Rule. Each annotated erroneous span is evaluated in an independent replay, conditioned on the original image, question, and response prefix preceding that span.

The prefix is fixed to the original response, while tokens from the beginning of the target span onward are generated by the model. Intervention starts with the decoding step that predicts the first token of the span.

After each intervened step, we compare the generated token with the original token at the corresponding position. If they match and the span has not ended, intervention continues at the next step. At the first mismatch, the newly generated token is retained and intervention is disabled from the next step onward. If all generated tokens match the original span, intervention ends after its last token. In either case, generation continues without further attention intervention until the response is complete. Labels from the original response are not transferred to the divergent continuation, and no later annotated spans are intervened on within the same replay. This stopping rule confines the use of the original annotation to decoding steps whose preceding token sequence still matches the original response. All replays use greedy decoding with identical settings across conditions. The no-intervention control ($\lambda = 0$) is evaluated using the same replay protocol and yields a recovery rate of 0%.

Span Statistics and Recovery Rate. The number of erroneous spans per response is small in our diagnostic set, with all 200 responses containing at least one visually-oriented and one text-oriented erroneous span. For the visually-oriented intervention set, 186 of the 200 responses contain one erroneous span and the remaining 14 contain two. For the text-oriented intervention set, 174 responses contain one erroneous span and the remaining 26 contain two. Thus, 93% and 87% of the responses in the visual and textual intervention sets, respectively, contain only a single erroneous span, while the remaining responses contain at most two.

Recovery rate is computed at the sample level. For each intervention condition and strength, a sample is counted as recovered if at least one independent span-level replay yields a correct final answer, and each sample is counted only once. The same aggregation rule is used for all control conditions. Importantly, replay attempts are matched across intervention conditions for each response; responses containing two annotated spans therefore receive the same number of replay attempts under Ours and the corresponding control conditions. Consequently, differences between intervention strategies are not driven by unequal replay budgets across conditions.

Causal Intervention Controls. We compare the role-aware intervention (Ours) with three controls that vary the intervention location or target modality using the 200 erroneous rollouts we annotated. No-label retains the intervention locations used by Ours but randomly selects the modality to enhance, without using the annotated functional roles. Random instead randomizes the intervention locations while retaining the designated target modality: visual attention in the visually-oriented intervention experiment and textual attention in the text-oriented intervention experiment. Wrong-Role retains the annotated intervention locations but enhances the modality opposite to the annotated role. We evaluate $\lambda \in \{0.1, 0.2, \ldots, 0.8\}$. For each condition and λ, recovery rate is the fraction of evaluated samples for which at least one independent replay yields a correct final answer, counting each sample only once.

Table 9: Controls for intervention on visually-oriented erroneous tokens.

P_λ	Ours	No-label	Random	Wrong-Role
0.0	0.0%	0.0%	0.0%	0.0%
0.1	7.0%	4.0%	5.5%	4.0%
0.2	15.5%	16.0%	11.5%	7.5%
0.3	13.0%	14.5%	10.5%	9.5%
0.4	21.5%	13.0%	15.0%	8.0%
0.5	32.5%	11.5%	20.0%	6.5%
0.6	23.5%	7.0%	14.5%	6.5%
0.7	28.0%	10.5%	14.5%	0.5%
0.8	11.0%	2.0%	8.5%	0.5%

Table 10: Controls for intervention on text-oriented erroneous tokens.

R_λ	**Ours**	**No-label**	**Random**	**Wrong-Role**
0.0	0.0%	0.0%	0.0%	0.0%
0.1	14.5%	11.0%	9.0%	7.5%
0.2	19.0%	15.0%	12.5%	4.0%
0.3	26.0%	21.5%	19.5%	8.5%
0.4	23.0%	13.5%	18.0%	9.0%
0.5	21.5%	17.0%	19.0%	7.5%
0.6	26.5%	14.0%	18.5%	4.0%
0.7	24.5%	7.5%	17.0%	2.5%
0.8	16.0%	1.5%	14.0%	1.0%

Tables 9 and 10 show that the role-aware intervention achieves higher recovery rates than the controls across most tested strengths. Wrong-Role generally yields the lowest recovery rates. Random usually performs below Ours, suggesting that the intervention location matters even when the target modality is retained. No-label remains competitive at small strengths but performs less well at larger strengths. Together, these comparisons support the benefit of jointly selecting where to intervene and which modality to enhance based on the annotated functional roles.

I Additional Analysis of Token Role Assignment

Alternative Metrics for Token Role Assignment. To isolate the effect of the role-assignment metric, we replace the Fisher–Rao score v_t^m with an alternative score g_t^m, while retaining the same role-assignment rule and downstream advantage reweighting. The Entropy variant uses the entropy of the within-modality attention distribution:

$$g_t^m = -\sum_{j \in \mathcal{I}_m} p_{t,j}^m \log p_{t,j}^m.$$

The KL variant measures the divergence of the current within-modality attention distribution from the preceding one:

$$g_t^m = D_{\mathrm{KL}}\big(p_t^m \,\|\, p_{t-1}^m\big) = \sum_{j \in \mathcal{I}_m} p_{t,j}^m \log \frac{p_{t,j}^m}{p_{t-1,j}^m}.$$

For the textual modality, the two distributions are restricted to their shared prefix and renormalized before computing KL divergence, following the Fisher–Rao comparison. Both variants use Eq. (8) with v_t^m replaced by g_t^m and $\tau = 0.05$; tokens satisfying neither condition remain neutral.

Evaluation of Fisher–Rao Role Assignment. We evaluate how well different attention-based measurements recover the human functional-role annotations introduced in Appendix F. We compare Fisher–Rao distance with Attention, Cosine Similarity, Entropy, and KL Divergence. Attention uses modality-specific attention mass, while Entropy measures the concentration of the within-modality attention distribution at each step. Cosine Similarity and KL Divergence compare attention distributions between consecutive steps. Fisher–Rao measures the angular distance between their square-root representations, providing a symmetric characterization of temporal redistribution within each modality.

For each measurement, we report three complementary statistics. *Coverage* denotes the proportion of human-annotated tokens that receive an explicit visual or textual role assignment. *Accuracy* is computed over these non-neutral predictions, with neutral assignments excluded from the denominator. Because the annotated set contains unequal numbers of visually- and text-oriented tokens, we additionally report *balanced accuracy*, computed as the mean of the per-class recalls over the non-neutral subset. Role assignments are obtained using the default stability margin $\tau = 0.05$.

Table 11: Agreement between attention-based role assignments and human token-role annotations.

Measurement	Coverage (%)	Accuracy (%)	Balanced Acc. (%)
Attention	26.2	60.5	55.3
Cosine Similarity	18.1	65.0	69.0
Entropy	24.7	67.0	56.3
KL Divergence	28.2	71.5	**73.5**
Fisher–Rao (Ours)	26.3	**74.0**	72.8

As shown in Table 11, Fisher–Rao achieves the highest overall agreement with the human annotations, reaching 74.0% accuracy on its non-neutral predictions. Its balanced accuracy (72.8%) is also comparable to that of KL divergence (73.5%), indicating that the higher overall accuracy is not obtained by strongly favoring one functional role. Together with its comparable prediction coverage, these results support Fisher–Rao as an effective empirical proxy for the functional roles used by DyCo-RL.

Distribution of Assigned Token Roles. We analyze 1,000 rollouts randomly sampled from GRPO+DyCo-RL training on ThinkLite-hard-11K, using the training stability margin $\tau = 0.05$. Table 12 summarizes the resulting token-role distribution. Overall, 29.7% of tokens are assigned visually-oriented or text-oriented roles and selected for role-conditioned advantage reweighting, while the remaining 70.3% retain their original sequence-level advantages. Neutral assignments indicate insufficient separation between the visual and textual attention changes under the threshold τ, rather than an absence of visual or textual processing.

Table 12: Distribution of token roles assigned by the Fisher–Rao criterion with $\tau = 0.05$.

Token role	Proportion (%)
Visually-oriented ($\mathcal{D}^{\mathrm{vis}}$)	10.9
Text-oriented ($\mathcal{D}^{\mathrm{txt}}$)	18.8
Neutral	70.3

Sensitivity to the Stability Margin τ. The stability margin τ is introduced to prevent small and potentially ambiguous differences between visual and textual attention transitions from forcing a binary role assignment. A token is assigned a functional role only when one modality transition sufficiently dominates the other; otherwise, it is treated as neutral, excluded from the role-aware normalization set, and retains the original sequence-level advantage.

Importantly, this neutral region also prevents tokens with no clear modality dominance, including potentially hybrid tokens that jointly depend on visual grounding and textual context, from being artificially pushed toward either modality. Thus, DyCo-RL does not impose a mutually exclusive visual/textual partition over all reasoning tokens; the binary role distinction is applied only to tokens exhibiting sufficiently confident modality-specific dynamics, while ambiguous cross-modal tokens remain unchanged. The sensitivity experiment therefore examines not only which value produces the highest score, but also whether DyCo-RL remains effective under different degrees of role-assignment conservativeness. We evaluate $\tau \in \{0, 0.05, 0.1\}$ on both Qwen2.5-VL-3B and Qwen2.5-VL-7B while keeping the remaining configuration fixed.

Table 13: Sensitivity analysis of the token role assignment stability margin τ.

Model	τ	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
Qwen2.5-VL-3B	GRPO	23.0	21.4	34.4	72.8	39.8	60.8	79.9	47.4
	0	25.1	20.7	35.6	74.6	**41.8**	61.8	80.6	48.6
	0.05	**25.9**	22.4	**36.0**	**76.2**	41.2	**62.4**	**80.7**	**49.3**
	0.1	25.2	**22.7**	35.2	75.8	40.7	61.5	80.1	48.7
Qwen2.5-VL-7B	GRPO	38.7	26.6	47.4	78.0	46.1	69.2	83.3	55.6
	0	40.0	28.0	49.3	81.8	48.1	70.0	83.3	57.2
	0.05	**40.5**	**29.3**	**49.5**	**83.8**	**48.5**	**71.6**	**84.9**	**58.3**
	0.1	40.3	27.3	49.2	82.9	48.1	69.2	84.6	57.4

Table 14: Ablation study on reweight strength α. Bold formatting indicates the highest performance in each column.

Method	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
GRPO Baseline ($\alpha = 0$)	23.0	21.4	34.4	72.8	39.8	60.8	79.9	47.4
$\alpha = 0.1$	25.7	23.4	34.5	74.3	39.8	**63.4**	80.1	48.7
DyCo-RL ($\alpha = 0.2$)	**25.9**	22.4	36.0	**76.2**	**41.2**	62.4	80.7	**49.3**
$\alpha = 0.3$	25.0	**24.7**	**36.4**	75.1	40.3	59.4	**81.7**	48.9
$\alpha = 0.4$	23.1	19.4	32.4	74.9	37.4	59.9	79.0	46.6
$\alpha = 0.5$	21.0	17.8	31.5	71.1	35.3	58.6	78.2	44.8

Table 13 shows that DyCo-RL improves over GRPO under every tested value of τ on both model sizes. Its effectiveness therefore does not depend on a narrowly tuned threshold. Among the tested values, $\tau = 0.05$ achieves the highest average performance for both Qwen2.5-VL-3B and Qwen2.5-VL-7B. This result suggests that a moderate margin provides an effective balance: it filters uncertain modality transitions while retaining tokens with sufficiently informative functional preferences. Notably, $\tau = 0.05$ achieves higher average performance than $\tau = 0$ on both model sizes, suggesting that forcing every token into a visual or textual role may not be necessary. Preserving a neutral region can help avoid assigning modality-specific roles to tokens whose multimodal function cannot be confidently reduced to a single dominant modality.

J Additional Ablation Studies

Ablation on Reweight Strength (α). Table 14 presents the ablation study on the reweight strength α, which controls the intensity of modality-aware advantage adjustment during RL training. Overall, moderate reweighting strengths improve average visual reasoning performance over the unweighted baseline ($\alpha = 0$). Specifically, $\alpha = 0.2$ achieves the highest average performance, while $\alpha = 0.1$ and $\alpha = 0.3$ obtain the best results on several individual benchmarks. This suggests that moderate reweighting can effectively incorporate modality-aware guidance while preserving stable reasoning dynamics.

Conversely, when α is too small (e.g., $\alpha = 0.1$), the modulation is relatively limited. At the other extreme, larger values ($\alpha \geq 0.4$) degrade average performance, with $\alpha = 0.4$ already falling below the GRPO baseline. We hypothesize that overly aggressive reweighting may disrupt the model's intrinsic coordination and lead to less effective optimization. Overall, the results favor moderate reweighting strengths, while also revealing sensitivity to excessively large α values.

Advantage Reweighting vs. Reward Shaping. As shown in Table 15, we further contrast our token-level advantage reweighting with a straightforward trajectory-level reward shaping baseline. We aggregate the token-level alignment scores over the response as $\bar{s} = \frac{1}{T}\sum_{t=1}^{T} s_t$ and augment the accuracy reward:

$$R' = R_{\text{acc}} + \beta \bar{s}, \tag{12}$$

where $\beta = 0.2$ is the reward shaping coefficient. All other reward components and training settings remain unchanged.

While reward shaping improves the average performance over the vanilla GRPO baseline and performs favorably on some individual benchmarks, its overall performance remains below advantage reweighting. We hypothesize that this difference arises partly from their distinct interaction with GRPO optimization. Reward shaping modifies the trajectory-level reward before group normalization and consequently changes the resulting group-relative advantage estimates. Advantage reweighting, by contrast, introduces token-level alignment guidance after the baseline computation, providing more fine-grained modulation without directly altering the trajectory-level reward.

Scaling the Number of Training Rollouts. The number of rollouts per prompt, G, controls the trade-off between training cost and performance. We evaluate DyCo-RL with $G \in \{4, 8, 16\}$ while holding all other hyperparameters constant. As shown in Table 16, its average score increases from 49.3 to 49.9 and 50.7, respectively. These results show that DyCo-RL benefits from larger rollout

Table 15: Comparison between Advantage Reweighting and Reward Shaping. Bold formatting indicates the highest performance in each column.

Method	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
GRPO Baseline	23.0	21.4	34.4	72.8	39.8	60.8	79.9	47.4
DyCo-RL (Reward Shaping)	24.4	19.7	34.5	75.5	37.6	**62.4**	80.3	47.8
DyCo-RL (Adv. Reweighting)	**25.9**	**22.4**	**36.0**	**76.2**	**41.2**	**62.4**	**80.7**	**49.3**

groups within the evaluated range, although they require additional computation for rollout generation and policy optimization. We adopt $G = 4$ as the default configuration for resource efficiency.

Table 16: Ablation Study on the Number of Training Rollouts (G).

Method	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
DyCo-RL ($G = 4$, Default)	25.9	22.4	**36.0**	76.2	41.2	62.4	80.7	49.3
DyCo-RL ($G = 8$)	27.1	23.0	35.3	77.6	41.4	62.7	82.0	49.9
DyCo-RL ($G = 16$)	**28.0**	**23.4**	**36.0**	**79.3**	**42.3**	**63.6**	**82.5**	**50.7**

Centered Advantage Reweighting. Our formulation applies the role-aware factor $1 + \alpha\,\mathrm{sgn}(\hat{A})w_t$ to non-neutral tokens. Since w_t averages to one over $\mathcal{D}$, this factor introduces a common shift in addition to relative token-level reweighting. To examine whether the improvement persists after removing this common component, we evaluate a centered variant:

$$\tilde{A}_t^{\mathrm{center}} = \begin{cases} \big[1 + \alpha\,\mathrm{sgn}(\hat{A})(w_t - 1)\big]\hat{A}, & t \in \mathcal{D}, \\ \hat{A}, & t \notin \mathcal{D}. \end{cases} \tag{13}$$

Because $\sum_{t\in\mathcal{D}}(w_t - 1) = 0$, the centered variant preserves the mean token-level advantage within each response, while retaining alignment-dependent differences between tokens. Neutral tokens remain unchanged. This preservation of the mean advantage does not imply preservation of the policy-gradient norm.

Table 17: Comparison between centered and original advantage reweighting on GRPO.

Method (Qwen2.5-VL-3B)	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
GRPO	23.0	21.4	34.4	72.8	39.8	60.8	79.9	47.4
GRPO + DyCo-RL (centered)	25.3	**23.4**	35.7	**76.2**	39.1	**62.4**	**80.7**	49.0
GRPO + DyCo-RL	**25.9**	22.4	**36.0**	**76.2**	**41.2**	**62.4**	**80.7**	**49.3**

As shown in Table 17, the centered variant also improves average performance over GRPO. This result suggests that the improvement does not depend solely on the common shift in token-level advantages and supports the benefit of alignment-guided relative reweighting.

K Multi-Seed Evaluation

Reinforcement-learning optimization can vary across random seeds. To examine whether the improvements of DyCo-RL are sensitive to training randomness, we conduct additional experiments with three training seeds, 42, 43, and 44, controlling training-data shuffling and rollout sampling. The evaluation seed and decoding settings are held fixed across runs. For each training seed, GRPO and GRPO + DyCo-RL are trained on Qwen2.5-VL-3B and Qwen2.5-VL-7B using otherwise identical training configurations.

Table 18 reports the mean and standard deviation across the three runs. DyCo-RL consistently improves over GRPO across all evaluated benchmarks and both model scales. In terms of average performance, DyCo-RL improves the score from 47.4 to 49.2 on Qwen2.5-VL-3B and from 55.8 to 58.1 on Qwen2.5-VL-7B. The standard deviations remain modest across repeated runs, suggesting that the observed gains are consistently preserved under the training variability examined here rather than being attributable to a particular random seed.

Table 18: Multi-seed evaluation. Each entry reports mean ± standard deviation over three matched random seeds.

Model	WeMath	MathVision	MathVerse	MMBench	LogicVista	HallusionBench	MME	Avg.
Qwen2.5-VL-3B								
GRPO	$23.0_{\pm 0.2}$	$21.3_{\pm 0.2}$	$34.4_{\pm 0.2}$	$72.8_{\pm 0.4}$	$39.9_{\pm 0.3}$	$60.8_{\pm 0.2}$	$79.8_{\pm 0.3}$	$47.4_{\pm 0.3}$
GRPO + DyCo-RL	$25.9_{\pm 0.3}$	$22.2_{\pm 0.4}$	$35.9_{\pm 0.2}$	$76.2_{\pm 0.6}$	$40.8_{\pm 0.4}$	$62.4_{\pm 0.2}$	$80.8_{\pm 0.3}$	$49.2_{\pm 0.2}$
Qwen2.5-VL-7B								
GRPO	$38.6_{\pm 0.2}$	$26.6_{\pm 0.4}$	$47.6_{\pm 0.2}$	$78.4_{\pm 0.5}$	$46.4_{\pm 0.5}$	$69.2_{\pm 0.3}$	$83.6_{\pm 0.4}$	$55.8_{\pm 0.3}$
GRPO + DyCo-RL	$40.3_{\pm 0.3}$	$29.0_{\pm 0.4}$	$49.5_{\pm 0.3}$	$83.5_{\pm 0.5}$	$48.1_{\pm 0.5}$	$71.7_{\pm 0.3}$	$84.8_{\pm 0.4}$	$58.1_{\pm 0.3}$

L Analysis of Computational Overhead

As detailed in Table 19, DyCo-RL reduces training throughput on Qwen2.5-VL-3B from 0.447 to 0.327 samples/sec, a decrease of 26.8%.

The overhead primarily arises from extracting and processing attention weights during training rollouts, including explicit attention materialization and modality-wise aggregation. Fisher–Rao distance computation and token role assignment introduce additional processing. Attention statistics are collected during autoregressive rollout generation, without a separate full-model forward pass after the rollout solely for attention extraction.

Table 19: Analysis of training throughput and computational overhead.

Method (Qwen2.5-VL-3B)	Throughput (samples/sec)	Δ Throughput
GRPO	0.447	–
GRPO + DyCo-RL	**0.327**	**-26.8%**

These additional operations are required only during training. At inference time, the trained model can use the standard backbone inference pipeline without attention-statistics extraction, role assignment, or advantage reweighting. DyCo-RL therefore introduces no additional trainable parameters or method-specific inference computation, while incurring the training overhead reported above.

M 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.

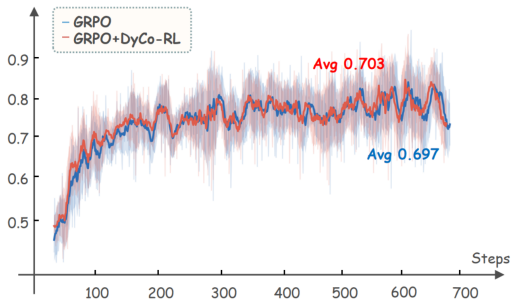


Figure 5: Training accuracy reward of GRPO and GRPO + DyCo-RL. Curves are smoothed using a rolling mean with a window size of 10.

As illustrated in Fig. 5, both methods show rapid initial increases in accuracy reward, followed by fluctuations around similar levels. DyCo-RL achieves a slightly higher average reward over the training period (0.703 versus 0.697), although the curves overlap and cross at multiple stages. This

comparison suggests a modest overall benefit of role-aware advantage reweighting for training reward optimization.

N Extended Generalization Analysis

A critical question regarding token-level advantage reweighting is whether the learned alignment signal overfits to the visual-mathematical distribution seen during training. To assess its out-of-domain robustness, we further evaluate DyCo-RL on A-OKVQA (Schwenk et al., 2022), RealWorldQA (xAI, 2024), MMStar (Chen et al., 2024), SEED-Bench (Li et al., 2023), ChartQA (Masry et al., 2022), and MMMU (Yue et al., 2024), covering commonsense visual understanding, real-world reasoning, general multimodal understanding, chart reasoning, and multidisciplinary expert-level reasoning.

Table 20: Out-of-domain generalization results. Bold formatting indicates the higher performance on each benchmark.

Method (Qwen2.5-VL-3B)	A-OKVQA	RealWorldQA	MMStar	SEED-Bench (IMG)	ChartQA (Test)	MMMU (Val)
GRPO	82.3	47.0	51.2	68.2	81.0	42.2
GRPO + DyCo-RL	**82.7**	**49.5**	**53.2**	**69.0**	**83.2**	**45.2**

As shown in Table 20, DyCo-RL consistently outperforms GRPO across all six benchmarks, demonstrating that the learned alignment strategy generalizes beyond the visual-mathematical training distribution to diverse multimodal reasoning and understanding tasks. These consistent gains across different domains and task formats further support the generalizability between domains of the proposed role-aware reweighting.

O 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:

Prompt Template

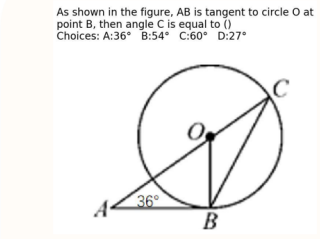
SYSTEM: You are a helpful assistant.

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}

The template above shows the textual component of the prompt, images are supplied through the model's standard multimodal processor and chat template.

P 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. Red text marks erroneous reasoning spans in the baseline responses, while green text marks correct reasoning spans in our model's responses. The selected examples illustrate how DyCo-RL can interleave visual evidence extraction and text-based reasoning, complementing the quantitative analyses in the main paper.



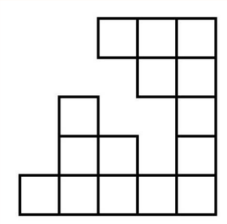
According to the question shown in the image, please directly answer the question and provide the correct option letter, e.g., A, B, C, D.

Baseline

<think>Since 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>

OURS:

<think>Since AB is tangent to circle O at point B, angle ABO is 90 degrees. Given that angle A is 36 degrees, angle AOB 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>



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?

Baseline

<think>\nThe 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</answer>

OURS:

<think>\nThe 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</answer>

Figure 6: Visualization Cases on Mathematical and Logical Reasoning.

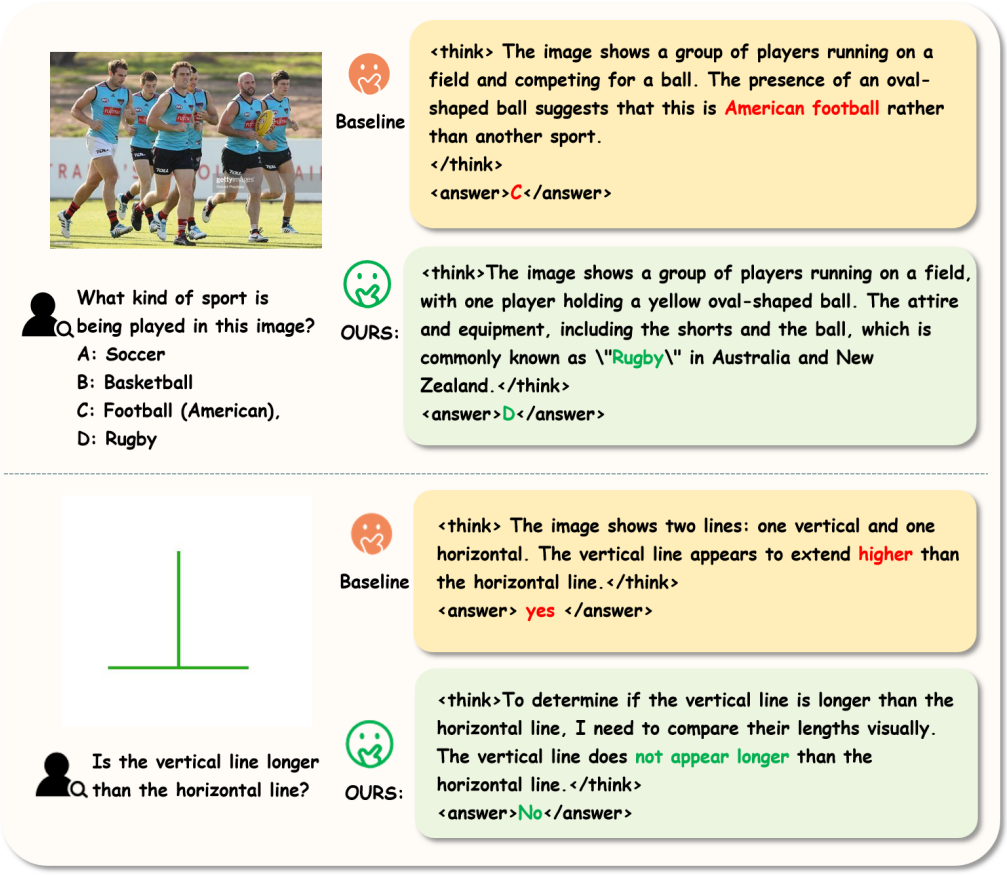


Figure 7: Visualization Cases on General Task and Hallucination Reasoning.