CREDIT WHERE IT'S DUE: CROSS-MODALITY CONNECTIVITY DRIVES PRECISE REINFORCEMENT LEARNING FOR MLLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) has significantly advanced the reasoning capabilities of Multimodal Large Language Models (MLLMs), yet how visual evidence is integrated during reasoning remains poorly understood. We explore multimodal RLVR through the lens of cross-modal attention connectivity and find that only a small fraction of tokens (approximately 15%) exhibit strong visual-textual coupling. These high-connectivity tokens act as anchors that ground reasoning in the image, while the majority follow linguistic patterns. During RLVR training, credit assignment naturally concentrates on these anchors, sharpening their visual grounding over time. Building on this insight, we propose Anchor-Token Reinforcement Learning (AT-RL), a lightweight framework that selectively reinforces high-connectivity tokens via graph-based clustering of attention topology. Evaluated across the Qwen2.5-VL series (3B–32B), AT-RL introduces only 1.2% overhead yet enables the 32B model to surpass the 72B-Instruct baseline on MathVista (80.2), with consistent gains observed across STEM, video and general tasks. Conversely, training solely on low-connectivity tokens causes severe degradation, supporting the view that precise credit assignment to visually grounded tokens is important for multimodal RL. Our results suggest that reasoning quality depends on the fidelity of cross-modal anchoring, rather than token quantity alone. These findings provide an operational account of how visual evidence is selected and reinforced during multimodal RLVR.

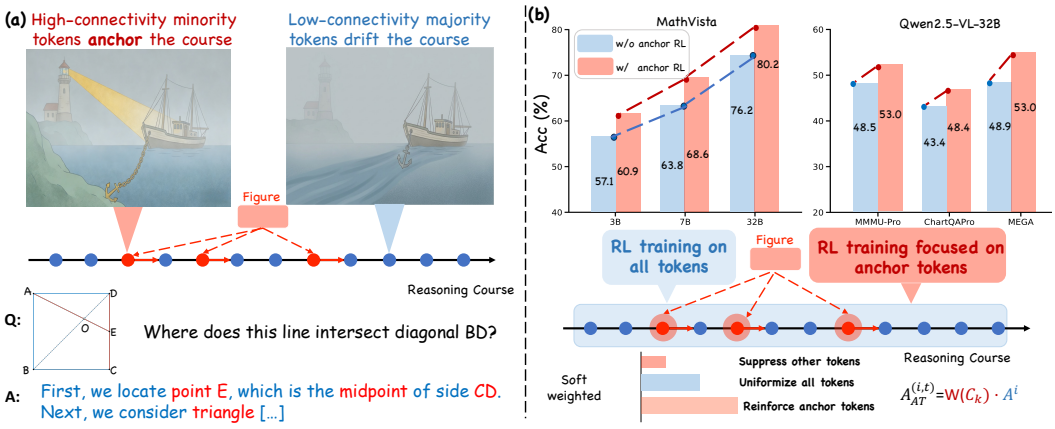

Figure 1: (a) In multimodal CoTs, only a minority of tokens (approximately 15%) exhibit high cross-modal connectivity and act as "perceptual anchors" that ground reasoning in visual evidence, while the majority are low-connectivity tokens fulfilling linguistic structures. (b) RLVR using AT-RL to modulate reinforcement signals via cluster-based soft weighting yields substantial reasoning improvements that scale from 3B to 32B.

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.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have demonstrated significant proficiency in complex domains such as mathematics and programming. While pre-training provides foundational capabilities, post-training through Reinforcement Learning (RL) has become the primary methodology for eliciting structured Chain-of-Thought (CoT) reasoning. The success of models such as OpenAI o1 and DeepSeek-R1 highlights the effectiveness of RL in optimizing reasoning trajectories, proving that models can resolve sophisticated problems by exploring an expanded solution space. Central to this development is Reinforcement Learning with Verifiable Rewards (RLVR), where online algorithms such as GRPO, GSPO, SAPO, and DAPO serve as scalable frameworks for multimodal policy optimization. In these reasoning tasks, the model processes a heterogeneous input sequence composed of interleaved visual patches and linguistic tokens, yet it generates a response serialized as a purely textual sequence. Consequently, while multimodal RLVR has achieved impressive gains through outcome-based rewards, existing implementations indiscriminately broadcast credit across all tokens, failing to address the challenge of distinguishing which tokens genuinely incorporate visual evidence versus those that follow linguistic patterns.

In this paper, we analyze the underlying mechanisms of multimodal RLVR through an innovative lens of attention connectivity patterns, investigating how the connectivity, representing the aggregated attention weights between textual tokens and visual patches, impacts reasoning performance. We first point out that in the Chain-of-Thought (CoT) processes of MLLMs, the connectivity distribution exhibits a distinct pattern where the majority of tokens are generated with low connectivity, while a critical minority of tokens emerge with high connectivity. Through comparing the functional roles of these two groups, we observe that the tokens with high connectivity, which we define as *perceptual anchors*, function as the lighthouse that grounds the reasoning course in visual evidence, while the tokens with lower connectivity act as the engine that completes linguistic structures and maintains computational coherence, as depicted in Figure 1(a). In addition to qualitative analysis, we conduct controlled experiments by modulating the reinforcement weights based on connectivity levels during the policy gradient update. Quantitative results reveal that a soft weighting approach, which prioritizes high-connectivity anchors while retaining the contribution of other tokens, leads to improvements in reasoning performance across diverse benchmarks, whereas uniform weighting or hard truncation yields suboptimal gains, confirming the importance of credit assignment.

Furthermore, by analyzing the evolution of connectivity patterns during training, we find that our proposed AT-RL framework enables the reasoning model to undergo a structured optimization of its attention distribution. Unlike standard RLVR, which often inherits the base model topology without significant refinement, AT-RL progressively sharpens the connectivity of perceptual anchors to improve grounding precision, thereby distinguishing them from the linguistic-heavy tokens. The above observations highlight the critical role that credit assignment may play in bridging visual inputs and textual logic during training, provided that the reinforcement signal is properly guided.

Building upon the discovery of perceptual anchors, we propose *Anchor-Token Reinforcement Learning* (AT-RL), a lightweight and plug-and-play framework for connectivity-driven advantage assignment. Specifically, AT-RL constructs a token-level graph from aggregated attention weights and employs graph partitioning to group reasoning sequences into semantically coherent clusters. For each cluster, we derive a weight based on its inter-modal connectivity density to re-scale the sequence-level advantage into a fine-grained, token-level credit signal. Unlike binary gradient masking in methods like VPPO that risks losing essential reasoning context, our soft-weighting strategy concentrates learning signals on perceptual anchors while preserving other tokens with modulated weights to maintain linguistic and computational coherence. Extensive experiments across the Qwen2.5-VL series (3B to 32B) show that AT-RL consistently outperforms strong baselines including GRPO, DAPO, and VPPO. Notably, our 32B model trained with AT-RL surpasses the performance of the significantly larger 72B-Instruct model on multiple benchmarks. Furthermore, AT-RL exhibits robust generalization across diverse tasks including multimodal STEM reasoning, general VQA, and complex video reasoning. Conversely, we find that random weighting, reverse weighting, or training exclusively on other tokens leads to severe performance degradation, confirming that multimodal RLVR efficacy is primarily driven by the precise optimization of cross-modal anchors.

Finally, we conduct detailed ablation studies to investigate the impact of key hyperparameters and the underlying mechanics of AT-RL. We find that our connectivity-driven weighting formula naturally concentrates reinforcement credit on approximately 15% of the tokens representing the perceptual anchors, while dynamically scaling the advantage of all other tokens based on their semantic clusters. This continuous weighting scheme ensures that the learning signal is precisely focused on visual-linguistic alignment without compromising the linguistic coherence of the reasoning chain. Notably, a single set of hyperparameters including clustering scales and connectivity thresholds generalizes across all evaluated benchmarks ranging from STEM reasoning to VideoMMMU, which underscores the intrinsic stability of the framework. Profiling results further show that the entire AT-RL module is remarkably efficient, introducing a negligible computational overhead of approximately 1.2% per training iteration. These findings confirm that the efficacy of multimodal RLVR is primarily driven by the precise cluster-based modulation of anchors, a strategy that aligns reinforcement signals with the structural interdependencies of vision and language to achieve reasoning precision. In summary, our main contributions are as follows:

1. **Discovery of Perceptual Anchors:** We identify a structural dichotomy in the attention topology of multimodal reasoning, where a critical minority of tokens act as *perceptual anchors* that ground the reasoning process in visual evidence, while the remaining tokens fulfill linguistic roles.
2. **The AT-RL Framework:** We propose *Anchor-Token Reinforcement Learning*, a lightweight, plug-and-play framework that constructs a token-level graph from attention weights and employs graph partitioning to group sequences into semantically coherent clusters. By re-scaling sequence-level advantages into fine-grained credit signals based on connectivity density, AT-RL enables precise optimization while preserving essential linguistic context.
3. **Scalable Generalization and Efficiency:** We demonstrate that AT-RL exhibits robust generalization across diverse tasks—including multimodal STEM reasoning, general VQA, and complex video reasoning. Our results show consistent gains from 3B to 32B, with the 32B model surpassing Qwen2.5-VL-72B-Instruct at only 1.2% overhead.

2 Related Work

Multimodal reasoning bottlenecks. Recent work reveals that insufficient visual perception constitutes the core bottleneck limiting multimodal reasoning. MATH-Vision finds models significantly underperform in vision-requiring tasks Wang et al. (2024). MathVerse confirms models fail to genuinely comprehend visual information and rely on textual cues Zhang et al. (2024b). ErrorRadar shows models struggle to identify visual-logical inconsistencies Yan et al. (2024). We-Math 2.0 reveals verbose responses often mask deficient visual understanding Qiao et al. (2025). GeoPQA quantifies a geometric perception gap where humans achieve near-perfect recognition while models frequently fail Chen et al. (2025). BLINK-Twice shows models lack fine-grained visual analysis for complex reasoning Ye et al. (2025). VisualToolBench quantifies that 70-80% of failures originate from visual perception errors Guo et al. (2025). MMErroR validates models cannot accurately localize error sources in visual-linguistic contexts Shi et al. (2026). These studies establish visual perception deficiency as the primary constraint on multimodal reasoning capability.

Reinforcement learning for LLM. PPO established the standard for RLHF Schulman et al. (2017), while DPO reframed alignment as supervised learning Rafailov et al. (2024). GRPO introduced group-relative advantage estimation without value networks Shao et al. (2024). DAPO introduced dynamic sampling and clip-higher mechanisms Yu et al. (2025). GVPO enhanced convergence via group variance control Zhang et al. (2025c). TreeRPO extended optimization to tree-structured trajectories Yang et al. (2025). GSPO enforced sequence-level policy updates Zheng et al. (2025). GMPO adopted geometric-mean objectives Zhao et al. (2025b). GTPO resolved gradient conflicts through trajectory-based advantage estimation Simoni et al. (2025). Lambda-GRPO reinterpreted GRPO as an implicit process reward model Sullivan & Koller (2025). CE-GPPO controlled entropy via gradient-preserving clipping Su et al. (2025). SAPO introduced soft adaptive mechanisms for dynamic policy regulation Gao et al. (2025). Despite diversity, these methods uniformly broadcast credit across tokens, overlooking roles.

Fine-Grained Credit Assignment. Different tokens play heterogeneous roles during reasoning. TokenDPO and Critical Tokens validated token-level optimization in preference alignment Zeng et al.

(2024); Lin et al. (2025). Beyond the 80/20 Rule reveals few high-entropy tokens drive performance in text reasoning Wang et al. (2025b). GCPO reinforces structural key tokens in image generation Zhang et al. (2025a). StepGRPO introduces step-level rewards, though requiring high-quality chain-of-thought Zhang et al. (2025b). VPPO reinforces visually-dependent tokens through binary gradient masking Huang et al. (2025). Zoom-Zero targets video temporal localization, SketchVL focuses on chart understanding, and Video-KTR selectively reinforces tokens using modality-aware signals Shen et al. (2025); Huang et al. (2026); Wang et al. (2026). These methods demonstrate that aligning credit assignment with token functional roles effectively enhances model capabilities.

3 PRELIMINARIES

3.1 CROSS-MODAL ATTENTION AND DEBIASING

The cross-modal attention weight from a generated textual token at index i to an input visual token at index j in layer l and head h is defined as:

$$A_h^l[i,j] = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})}, \quad e_{ij} = \frac{(W^Q \mathbf{o}_i) \cdot (W^K \mathbf{o}_j)^\top}{\sqrt{d}}.$$

Here, $\mathbf{o}_i$ and $\mathbf{o}_j$ are the hidden states at positions i and j, W^Q and W^K are learnable projection matrices, and d is the dimensionality of the attention head. This computation follows the standard attention mechanism Vaswani et al. (2017). When i indexes a textual token and j a visual token, $A_h^l[i,j]$ quantifies the model's visual grounding at step i, representing the raw, uncalibrated dependency of the generated text on a visual patch.

"Attention bias" corresponds to systematic distortions in the attention distribution. The raw attention is contaminated by positional preferences and inflated scores for padding tokens, i.e., $A_{\text{observed}} = A_{\text{semantic}} + A_{\text{bias}}$. We introduce a parametric calibration function $\mathcal{B}$ to remove this bias, obtaining a purified weight $\tilde{A}_h^l[i,j] = \mathcal{B}(A_h^l[i,j]; \Theta)$ for fine-grained credit assignment, extending recent debiasing work Zhao et al. (2025a).

3.2 REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Reinforcement Learning with Verifiable Rewards (RLVR) optimizes policies with a benchmark-defined verifier. Depending on the task, the verifier may use deterministic parsing and exact matching or the benchmark's official semantic-answer evaluator. We write the reward abstractly as $R(y;q)$ for output y and question q; the binary form below is a convenient special case for exact-match tasks:

$$R(y;q) = \begin{cases} +1, & \text{if the official verifier accepts } y, \\ -1, & \text{otherwise.} \end{cases} \tag{1}$$

Proximal Policy Optimization (PPO) Schulman et al. (2017) serves as the foundational policy-gradient algorithm for this paradigm. It stabilizes updates by constraining the probability ratio $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ within a clipped range $[1-\epsilon, 1+\epsilon]$. The objective is formulated as:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) A_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A_t \right) \right] \tag{2}$$

where A_t is the advantage typically estimated by an additional value network.

Group Relative Policy Optimization (GRPO) Shao et al. (2024) adapts the PPO framework to the RLVR setting by substituting the value network with group-normalized rewards. For a group of G responses $\{o_1, \ldots, o_G\}$ generated from a prompt q, the sequence-level advantage $A^{(i)}$ is computed as:

$$A^{(i)} = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^G)}{\text{std}(\{R_j\}_{j=1}^G)} \tag{3}$$

The final objective applies the PPO clipped surrogate loss sequence-wise and incorporates a KL term:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E} \left[\min \left(r_t(\theta) A, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A \right) \right] - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \tag{4}$$

Subsequent advancements, including DAPO Yu et al. (2025), GSPO Zheng et al. (2025), and SAPO Gao et al. (2025), build upon this framework to further enhance training stability and convergence. Our proposed AT-RL leverages these engines as a base.

4 ANCHOR-TOKEN REINFORCEMENT LEARNING (AT-RL)

We introduce **AT-RL** (**A**nchor-**T**oken **R**einforcement **L**earning), a policy optimization framework that resolves the uniform-credit limitation of existing RLVR algorithms. By leveraging intrinsic cross-modal attention, AT-RL concentrates reinforcement signals on critical tokens that ground reasoning in visual evidence, ensuring that gradients are prioritized for tokens essential to task success. As illustrated in Fig. 2, our method unfolds in three stages: (i) locating perceptual anchors via connectivity quantification, (ii) grouping tokens into coherent clusters through graph-based partitioning, and (iii) reinforcing credit assignment.

4.1 LOCATING PERCEPTUAL ANCHORS

The first stage of AT-RL is to extract internal signals from the Multimodal Large Language Model (MLLM) to identify the functional roles of tokens. Given a generated reasoning sequence $S = \{t_1, \ldots, t_T\}$ and the input visual patches $\mathcal{V}$, we extract the cross-modal attention matrices $\mathbf{A}_h^l \in \mathbb{R}^{T \times |\mathcal{V}|}$ from the top L transformer layers and H attention heads. Row i corresponds to generated token t_i and column j to visual patch v_j. To obtain a stable representation of the model's semantic footprint, we aggregate these matrices across both layers and heads:

$$\bar{\mathbf{A}} = \frac{1}{LH} \sum_{l=1}^{L} \sum_{h=1}^{H} \mathbf{A}_h^l. \tag{5}$$

Bias Correction. As discussed in Section 3.1, raw attention weights are often contaminated by structural artifacts, such as "attention sinks" at initial tokens or positional preferences, which do not reflect genuine semantic grounding. To isolate meaningful signals, we apply the debiasing operator $\mathcal{B}(\cdot)$ by normalizing each column j of the aggregated matrix with a mean-normalized bias curve b_j:

$$\tilde{\mathbf{A}}_{i,j} = \frac{\bar{\mathbf{A}}_{i,j}}{b_j}. \tag{6}$$

The curve b_j is indexed by visual-patch position ($j = 1, \ldots, |\mathcal{V}|$) and parameterized by a combination of exponential decay and periodic terms to capture attention-sink and patch-flattening biases (see Appendix B). This yields the calibrated attention matrix $\tilde{\mathbf{A}}$ used for subsequent credit assignment.

Connectivity Density. Based on the calibrated matrix $\tilde{\mathbf{A}}$, we define the **Connectivity Density** C_i for each generated textual token t_i. This metric quantifies the intensity with which a token extracts visual evidence during generation:

$$C_i = \sum_{j \in \mathcal{V}} \tilde{\mathbf{A}}_{i,j}. \tag{7}$$

Empirical observations across various reasoning tasks reveal a distinct structural dichotomy in multimodal traces. A critical minority of tokens—the **Perceptual Anchors**—exhibit disproportionately high connectivity density, typically corresponding to key observations or evidence extraction from the image. In contrast, the vast majority of tokens, which we refer to as **Other tokens**, possess C_i values near zero, as they primarily serve to maintain linguistic coherence and logical transitions. This heterogeneity suggests that the uniform-credit assumption is physically inaccurate, providing the foundation for our credit assignment.

4.2 GRAPH-BASED TOKEN GROUPING

While connectivity density C_i provides a per-token measure of grounding, individual token signals can be noisy or fragmented. To identify coherent reasoning steps, we group tokens into semantic clusters by representing the sequence as a graph and applying topological partitioning.

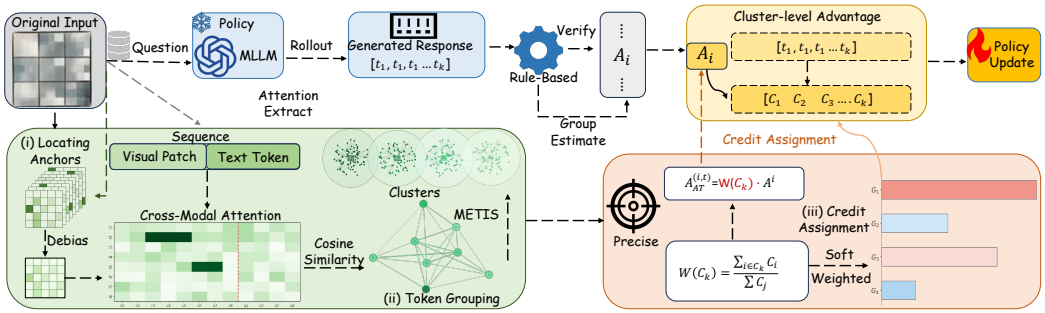

Figure 2: **Overview of the Anchor-Token Reinforcement Learning (AT-RL) framework.** Our method refines the reinforcement signal through a three-stage connectivity-aware analysis: **(i) Locating Anchors:** Extracting cross-modal attention footprints from the MLLM and applying debiasing to identify key perceptual tokens; **(ii) Token Grouping:** Constructing a functional-similarity graph based on attention similarity and partitioning it into semantic clusters $\{G_k\}$ via the METIS algorithm; **(iii) Credit Assignment:** Quantifying cluster-level importance $W(G_k)$ to perform soft-weighting on the base advantage A^i. The resulting modulated advantage $A_{AT}^{(i,t)}$ ensures that policy gradients are prioritized for tokens essential to visual grounding while preserving the logical coherence of the reasoning chain.

Graph Construction. We construct a weighted undirected graph $G = (V, E)$, where each vertex $v_i \in V$ corresponds to a token t_i in the reasoning sequence. The graph represents functional similarity between tokens, rather than the direction of information flow. The edge weight w_{ij} between two tokens t_i and t_j is defined by the cosine similarity of their calibrated attention footprints (rows of $\tilde{\mathbf{A}}$):

$$w_{ij} = \frac{\tilde{\mathbf{A}}_i \cdot \tilde{\mathbf{A}}_j}{\|\tilde{\mathbf{A}}_i\| \|\tilde{\mathbf{A}}_j\|}. \tag{8}$$

To ensure a sparse and semantically relevant graph, an edge is established only if $w_{ij} > \tau_{\text{sim}}$, where τ_{sim} is a similarity threshold. This graph captures the functional dependencies within the reasoning trace, where tokens with similar visual grounding and linguistic contexts are more densely connected.

Sequence Partitioning via METIS. We partition the graph G into K disjoint semantic clusters $\{G_1, \ldots, G_K\}$ using the METIS algorithm Karypis & Kumar (1998). METIS is chosen for its efficiency in minimizing the edge cut—the total weight of edges crossing cluster boundaries—thereby ensuring that tokens within the same cluster share maximum semantic affinity. The number of clusters K is set proportional to the sequence length T to adapt to varying reasoning complexities.

Structural Refinement and Propagation. To ensure the robustness of the identified clusters, we perform a two-step refinement process to handle attention noise and semantic continuity. First, we apply *centroid-based denoising* to suppress spurious connections that may bypass initial partitioning. For each cluster G_k, we compute its semantic centroid $\mathbf{c}_k$ in the attention space. Tokens that exhibit low cosine similarity to their respective centroid ($\cos(\tilde{\mathbf{A}}_i, \mathbf{c}_k) < \tau_{\text{cen}}$) have their credit weight attenuated by a factor α, thereby purifying the cluster's signal. Second, to preserve the continuity of reasoning steps that transcend discrete cluster boundaries, we execute a *topological propagation* step. We identify central nodes with high graph importance within the top-q percentile (default $q = 0.15$) and propagate their reinforcement influence to their R-nearest neighbors. This mechanism effectively recovers transition tokens (Other tokens) that are topologically vital to the Perceptual Anchors, resulting in a set of semantically dense clusters that serve as the foundational units for fine-grained credit assignment.

4.3 Reinforced Credit Assignment

The final stage of AT-RL translates the topological structures identified above into differential reinforcement signals. The core of this mechanism lies in modulating the advantage signal at the cluster level, thereby resolving the limitations of uniform credit assignment in multimodal reasoning.

Quantifying Cluster Importance. To measure the contribution of each reasoning step to visual perception, we define the cluster mass $W(G_k)$. This mass is calculated as the proportion of the sequence's total connectivity density that is concentrated within a specific cluster G_k:

$$W(G_k) = \frac{\sum_{i \in G_k} C_i}{\sum_{j=1}^{T} C_j}. \tag{9}$$

This quantitative metric reflects the "perceptual load" across different segments of the reasoning trace. Clusters containing Perceptual Anchors naturally receive higher mass, while those primarily serving linguistic transitions are assigned lower weights.

Modulated Advantage and Objective. AT-RL does not alter the fundamental loss structure of the underlying RL engine; instead, it acts as an advantage modulation layer. As illustrated in Fig. 1(b), given the original sequence-level advantage $A^{(i)}$ generated by a base engine (e.g., GRPO), we map it to a token-level modulated advantage. We normalize by cluster size so that the mean token multiplier remains one and the base optimizer's gradient scale is preserved:

$$\omega_t = \frac{TW(G_k)}{|G_k|}, \qquad A_{\text{AT}}^{(i,t)} = \omega_t A^{(i)}, \quad \forall t \in G_k. \tag{10}$$

When connectivity is nearly flat, this size-normalized weighting approaches uniform credit; when it is concentrated, it prioritizes clusters with higher visual connectivity while retaining non-anchor tokens. By reshaping the advantage signal based on semantic roles, we integrate this modulated signal into the standard policy gradient objective. The final AT-RL optimization function is formulated as follows:

$$\mathcal{L}_{\text{AT-RL}}(\theta) = \mathbb{E}\left[\min\left(r_t(\theta)A_{\text{AT}}, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_{\text{AT}}\right)\right] - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \tag{11}$$

By substituting the uniform $A^{(i)}$ with the fine-grained $A_{\text{AT}}^{(i,t)}$, AT-RL ensures that policy gradients are prioritized for tokens essential to task success. This soft-weighting scheme reinforces visual grounding while preserving the linguistic integrity of the reasoning chain. Unlike binary masking, this approach maintains low-weight gradients for non-anchor tokens, ensuring that the model does not lose its ability to generate logical and coherent transitions. Because the modulation occurs at the advantage level, AT-RL is inherently plug-and-play, allowing for seamless integration into existing engines like PPO, GRPO, or SAPO without requiring modifications to the policy architecture.

5 Experiments

5.1 Experiment Setup

Models. We evaluate AT-RL across the Qwen2.5-VL model family, spanning three parameter scales: 3B, 7B, and 32B. All models use the Instruct variant as the base policy.

Training Data. Following established practice Wang et al. (2025a), we train on a curated subset of geometry-focused reasoning data. Specifically, we use the Geometry-3K dataset Lu et al. (2021), supplemented with GeoQA Chen et al. (2022), yielding approximately 3,000 training samples with verifiable answers.

Baselines. We compare AT-RL against the following baselines: (1) Zero-shot: direct inference without any RL training; (2) GRPO Shao et al. (2024): the standard group-relative policy optimization with uniform credit; (3) DAPO Yu et al. (2025): dynamic sampling with clip-higher mechanism; (4) GSPO Zheng et al. (2025): group sequence policy optimization; (5) VPPO Huang et al. (2025): visual perception-based token masking; (6) FT-RL: a reverse-connectivity variant that exclusively reinforces low-connectivity (non-anchor) tokens, serving as a controlled ablation.

Evaluation Benchmarks. We evaluate on six mathematical reasoning benchmarks: MathVerse Zhang et al. (2024b), MathVista Lu et al. (2024), MathVision Wang et al. (2024), GeoQA Chen et al. (2022), and Inter-GPS Lu et al. (2021). For broader evaluation, we additionally report on MMMU-Pro Yue et al. (2025), MMMU Yue et al. (2024), EMMA Hao et al. (2025), and video reasoning benchmarks.

Table 1: Main results of Qwen2.5-VL-7B-Instruct with different RLVR training methods. AT-RL is a universal plug-and-play enhancement that consistently improves various base RL algorithms. Arrow values represent absolute changes relative to the *Zero-shot baseline*.

Method	$\text{GeoQA}_{\text{test}}$	MathVerse	MathVision	MathVista	WeMath	**Average**
Zero-shot	44.05	39.62	22.70	61.10	57.59	45.01
Uniform Credit Baselines						
GRPO	46.86↑2.81	41.81↑2.19	24.37↑1.67	63.80↑2.70	59.71↑2.12	47.31↑2.30
DAPO	47.54↑3.49	42.61↑2.99	25.36↑2.66	64.30↑3.20	60.41↑2.82	48.04↑3.03
GSPO	47.64↑3.59	44.27↑4.65	25.20↑2.50	66.40↑5.30	59.15↑1.56	48.53↑3.52
SAPO	49.46↑5.41	44.78↑5.16	27.63↑4.93	66.00↑4.90	61.82↑4.23	49.94↑4.93
Fine-Grained Credit Baselines						
StepGRPO	41.12↓2.93	35.34↓4.28	19.80↓2.90	59.70↓1.40	55.23↓2.36	42.24↓2.77
FT-RL	44.19↑0.14	39.14↓0.48	22.37↓0.33	61.20↑0.10	57.35↓0.24	44.85↓0.16
VPPO	48.89↑4.84	45.12↑5.50	28.03↑5.33	65.80↑4.70	60.97↑3.38	49.76↑4.75
Anchor-RL (Ours - Universal Plug-and-Play)						
GRPO + AT-RL	49.24↑5.19	45.98↑6.36	27.97↑5.27	66.70↑5.60	62.73↑5.14	50.52↑5.51
DAPO + AT-RL	50.13↑6.08	47.63↑8.01	28.73↑6.03	68.10↑7.00	63.56↑5.97	51.63↑6.62
GSPO + AT-RL	50.74↑6.69	47.46↑7.84	29.32↑6.62	68.10↑7.00	64.46↑6.87	52.02↑7.01
SAPO + AT-RL	**52.77**↑8.72	**48.34**↑8.72	**30.33**↑7.63	**68.60**↑7.50	**66.23**↑8.64	**53.25**↑8.24

Table 2: Scaling analysis of AT-RL comparing 3B and 7B (Expanded Data) models. Arrow values represent absolute changes relative to the *Zero-shot baseline*.

Method	$\text{GeoQA}_{\text{test}}$	MathVerse	MathVision	MathVista	WeMath	**Average**
Qwen2.5-VL-7B (Expanded Training Data)						
Zero-shot	44.05	39.62	22.70	61.10	57.59	45.01
GRPO	45.80↑1.75	41.30↑1.68	24.10↑1.40	62.90↑1.80	59.20↑1.61	46.66↑1.65
DAPO	46.50↑2.45	42.20↑2.58	24.80↑2.10	63.50↑2.40	59.80↑2.21	47.36↑2.35
DAPO + AT-RL (Ours)	**48.55**↑4.50	**45.30**↑5.68	**27.40**↑4.70	**66.50**↑5.40	**62.10**↑4.51	**49.97**↑4.96
Qwen2.5-VL-3B						
Zero-shot	38.72	32.97	17.96	55.20	51.78	39.33
GRPO	40.65↑1.93	34.92↑1.95	19.65↑1.69	57.15↑1.95	53.62↑1.84	41.20↑1.87
DAPO	41.70↑2.98	35.95↑2.98	20.60↑2.64	58.15↑2.95	54.65↑2.87	42.21↑2.88
DAPO + AT-RL (Ours)	**43.85**↑5.13	**38.90**↑5.93	**22.80**↑4.84	**60.90**↑5.70	**58.30**↑6.52	**44.95**↑5.62

5.2 Main Results

Overall Performance. As shown in Table 1, AT-RL significantly improves the multimodal reasoning capabilities across all benchmarks. When combined with SAPO, our method achieves an average accuracy of 53.25%, representing a substantial +8.24pp gain over the zero-shot baseline. Notably, AT-RL acts as a universal plug-and-play enhancement, improving GRPO by +3.21pp and DAPO by +3.59pp in average scores. These results demonstrate that reinforcing perceptual anchors provides a more effective gradient signal than treating all generated tokens equally (Uniform Credit) or assigning rewards blindly to each token (StepGRPO).

Scaling Results. Table 2 shows AT-RL provides consistent +5pp gains for both 3B and 7B models. Notably, the 3B model with AT-RL (44.95%) nearly matches the zero-shot performance of the 7B model (45.01%), demonstrating that reinforced grounding can compensate for smaller model capacity.

Foundation Model Comparison. Appendix I compares Qwen2.5-VL-32B + AT-RL against proprietary SOTA models. Our method achieves 80.2% on MathVista, significantly outperforming GPT-4o (60.0%) and even surpassing the larger Qwen2.5-VL-72B Instruct (77.8%). Across general VQA and Chart benchmarks, AT-RL maintains a competitive edge, proving it enhances general perceptual grounding beyond pure mathematical reasoning.

Table 3: Video reasoning performance of Qwen2.5-VL-7B on Video-R1 benchmark. Arrow values represent absolute percentage-point changes relative to the *Zero-shot baseline*.

	16 Frames			32 Frames			64 Frames		
Method	**VSI**	**VidMMMU**	**MMVU**	**VSI**	**VidMMMU**	**MMVU**	**VSI**	**VidMMMU**	**MMVU**
Zero-shot	27.7	47.8	59.2	30.1	48.1	60.0	32.5	48.5	61.2
SFT	31.5↑3.8	49.2↑1.4	60.8↑1.6	34.2↑4.1	50.5↑2.4	62.3↑2.3	36.8↑4.3	51.0↑2.5	63.5↑2.3
T-GRPO	33.8↑6.1	48.9↑1.1	62.1↑2.9	36.5↑6.4	50.2↑2.1	64.0↑4.0	39.2↑6.7	50.8↑2.3	65.0↑3.8
Ours	**36.4**↑8.7	**51.5**↑3.7	**64.8**↑5.6	**39.6**↑9.5	**53.4**↑5.3	**67.2**↑7.2	**44.3**↑11.8	**56.0**↑7.5	**70.2**↑9.0

5.3 Modality Generalization and Efficiency

Video Generalization. Table 3 demonstrates AT-RL's extensibility to temporal video understanding. Our variant consistently outperforms zero-shot, SFT, and T-GRPO. At 64 frames, our method achieves an average accuracy gain of +9.4pp over zero-shot. The gain on VSI-Bench (+11.8pp) is particularly notable, indicating that AT-RL effectively identifies critical temporal anchor tokens that ground reasoning across frames.

Computational Efficiency. Profiling on an NVIDIA A100 GPU reveals that the core backward and forward passes dominate execution time at 33.5% and 65.4%, respectively. In stark contrast, the entire AT-RL module (including attention processing, bias correction, clustering, and weight computation) introduces a negligible overhead of only 1.2%, confirming its practical scalability.

5.4 Ablation Studies

Table 4: Ablation on hard truncation strategies (7B on MathVerse).

Method	**MathVerse**
Zero-shot	39.62
GRPO (Uniform weighting)	41.48↑1.86
Random weighted	38.10↓1.52
Reverse weighted	36.80↓2.82
Hard truncation (top 5%)	41.12↑1.50
Hard truncation (top 10%)	42.12↑2.50
Hard truncation (top 15%)	42.58↑2.96
Hard truncation (top 20%)	42.42↑2.80
Hard truncation (top 30%)	41.95↑2.33
Hard truncation (top 50%)	41.68↑2.06
Hard truncation (top 75%)	41.30↑1.68
AT-RL (Soft weighting)	**45.27**↑5.65

Table 5: Ablation on bias correction (3B on MathVerse).

Method	**MathVerse**
Zero-shot	32.97
GRPO (Uniform)	35.10↑2.13
AT-RL w/o Bias Correction	34.20↑1.23
AT-RL w/ Bias Correction	**37.50**↑4.53

Table 6: Ablation on semantic grouping mechanisms.

Configuration	**MathVerse**	**MathVista**
Raw Attention Weighting	42.35	63.30
METIS Only	43.18	64.30
w/o Semantic Expansion	43.88	64.95
w/o Structural Refinement	43.78	64.90
Full AT-RL Pipeline	**44.60**	**65.60**

Ablation Results. Table 4 shows that soft weighting (45.27%) is superior to any hard truncation strategy, including the best-performing top 15% truncation (42.58%). Table 5 confirms that bias correction is critical, providing a +3.3pp gain over the uncorrected version. Finally, Table 6 proves that every step of our grouping pipeline—including semantic expansion and structural refinement—is essential for achieving maximum performance.

6 Conclusion

In this work, we introduced **AT-RL** (**A**nchor-**T**oken **R**einforcement **L**earning), a plug-and-play framework that addresses the uniform credit assignment limitation in existing RLVR methods by modulating advantage signals proportional to the semantic importance of perceptual anchors. By leveraging graph-based token grouping and topological refinement, AT-RL strategically concentrates reinforcement signals on perceptually critical content, significantly improving both early-stage visual grounding and overall reasoning accuracy. Extensive experiments across diverse multimodal benchmarks demonstrate consistent and substantial gains over strong baselines.

AI USE STATEMENT

The authors used a generative AI assistant for limited editorial support, LaTeX template conversion, and consistency checks. The authors reviewed and verified all scientific content, experimental results, citations, and final manuscript; no AI system generated data or experimental results.

REFERENCES

Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Deli Zhao, Anh Tuan Luu, and Yu Rong. GeoPQA: Bridging the visual perception gap in MLLMs for geometric reasoning, 2025. URL https://arxiv.org/abs/2509.17437.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning, 2022. URL https://arxiv.org/abs/2105.14517.

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.

Xingang Guo, Utkarsh Tyagi, Advait Gosai, Paula Vergara, Jayeon Park, Ernesto Gabriel Hernández Montoya, Chen Bo Calvin Zhang, Bin Hu, Yunzhong He, Bing Liu, and Rakshith Sharma Srinivasa. Beyond seeing: Evaluating multimodal LLMs on tool-enabled image perception, transformation, and reasoning, 2025. URL https://arxiv.org/abs/2510.12712.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs reason in multimodality? EMMA: An enhanced multimodal reasoning benchmark, 2025. URL https://arxiv.org/abs/2501.05444.

Muye Huang, Lingling Zhang, Yifei Li, Yaqiang Wu, and Jun Liu. SketchVL: Policy optimization via fine-grained credit assignment for chart understanding and more, 2026. URL https://arxiv.org/abs/2601.05688.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning, 2025. URL https://arxiv.org/abs/2510.09285.

George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs, 1998.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, 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, 2025. URL https://arxiv.org/abs/2411.19943.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning, 2021. URL https://arxiv.org/abs/2105.04165.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, Jie Wang, Chong Sun, Chen Li, and Honggang Zhang. We-Math 2.0: A versatile MathBook system for incentivizing visual mathematical reasoning, 2025. URL https://arxiv.org/abs/2508.10433.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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.

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-Zero: Reinforced coarse-to-fine video understanding via temporal zoom-in, 2025. URL https://arxiv.org/abs/2512.14273.

Yang Shi, Yifeng Xie, Minzhe Guo, Liangsi Lu, Mingxuan Huang, Jingchao Wang, Zhihong Zhu, Boyan Xu, and Zhiqi Huang. MMErroR: A benchmark for erroneous reasoning in vision-language models, 2026. URL https://arxiv.org/abs/2601.03331.

Marco Simoni, Aleksandar Fontana, Giulio Rossolini, Andrea Saracino, and Paolo Mori. GTPO: Stabilizing group relative policy optimization via gradient and entropy control, 2025. URL https://arxiv.org/abs/2508.03772.

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. CE-GPPO: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning, 2025. URL https://arxiv.org/abs/2509.20712.

Michael Sullivan and Alexander Koller. GRPO is secretly a process reward model, 2025. URL https://arxiv.org/abs/2509.21154.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning, 2025a. URL https://arxiv.org/abs/2504.08837.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset, 2024. URL https://arxiv.org/abs/2402.14804.

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, 2025b. URL https://arxiv.org/abs/2506.01939.

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-KTR: Reinforcing video reasoning via key token attribution, 2026. URL https://arxiv.org/abs/2601.19686.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, Aoxiao Zhong, Kun Wang, Hui Xiong, Philip S. Yu, Xuming Hu, and Qingsong Wen. ErrorRadar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection, 2024. URL https://arxiv.org/abs/2410.04509.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Xiaodan Liang, Yiwei Wang, and Jing Tang. TreeRPO: Tree relative policy optimization, 2025. URL https://arxiv.org/abs/2506.05183.

Junyan Ye, Dongzhi Jiang, Jun He, Baichuan Zhou, Zilong Huang, Zhiyuan Yan, Hongsheng Li, Conghui He, and Weijia Li. BLINK-Twice: You see, but do you observe? a reasoning benchmark on visual perception, 2025. URL https://arxiv.org/abs/2510.09361.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale, 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, 2024. URL https://arxiv.org/abs/2311.16502.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark, 2025. URL https://arxiv.org/abs/2409.02813.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization, 2024. URL https://arxiv.org/abs/2404.11999.

Guohui Zhang, Hu Yu, Xiaoxiao Ma, JingHao Zhang, Yaning Pan, Mingde Yao, Jie Xiao, Linjiang Huang, and Feng Zhao. Group critical-token policy optimization for autoregressive image generation, 2025a. URL https://arxiv.org/abs/2509.22485.

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, 2025b. URL https://arxiv.org/abs/2503.12937.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models, 2024a. URL https://arxiv.org/abs/2407.12772.

Kaichen Zhang, Yuzhong Hong, Junwei Bao, Hongfei Jiang, Yang Song, Dingqian Hong, and Hui Xiong. GVPO: Group variance policy optimization for large language model post-training, 2025c. URL https://arxiv.org/abs/2504.19599.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems?, 2024b. URL https://arxiv.org/abs/2403.14624.

Kai Zhao, Wubang Yuan, Yuchen Lin, Liting Ruan, Xiaofeng Lu, Deng-Ping Fan, Ming-Ming Cheng, and Dan Zeng. Attention debiasing for token pruning in vision language models, 2025a. URL https://arxiv.org/abs/2508.17807.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization, 2025b. URL https://arxiv.org/abs/2507.20673.

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 DISCUSSION

Discussion 1: Perceptual Anchors Emerge in Multimodal CoT. We analyze cross-modal connectivity by computing average attention weights each generated token receives from visual patches. Figure 3(a) shows connectivity density over 540 tokens from 100 MathVision samples is highly skewed: only **15.0%** exhibit high connectivity (above **0.0368**). High-connectivity tokens (Figure 3(b)) are dominated by action-oriented words like *"calculate"*, *"given"*, *"determine"*, serving as perceptual anchors bridging visual perception and logical derivation. Low-connectivity tokens (Figure 3(c)) consist of abstract terms like *"simplify"*, *"relation"*, maintaining linguistic coherence. These findings confirm cross-modal connectivity correlates with semantic function, providing the empirical foundation for AT-RL.

Discussion 2: Why Cluster Anchors? A Case Study. Individual token-level credit assignment is noisy due to surface-form variation and positional instability. We cluster high-connectivity tokens by semantic similarity to identify stable grounding signals. Figure 4(a) shows anchor importance is highly concentrated: clusters G_4 and G_0 dominate the distribution, justifying credit assignment on coherent units rather than isolated tokens. Figure 4(b) reveals these top clusters contain tokens consistently associated with visual grounding. G_4 includes function words like *"given"*, *"the"*, *"of"* that introduce diagram references. G_0 contains action terms like *"assume"*, *"divides"*, *"considering"* that trigger interpretation of spatial relations. This demonstrates clustering captures reliable grounding behaviors obscured at the single-token level.

Discussion 3: Limitations of RL and Complementarity with SFT. Figure 5 highlights a key limitation of RL: despite AT-RL's substantial reduction in **Visual Perception Errors (VPE)**, **Knowledge Deployment Errors (KDE)** remain largely unaffected. This is inherent to RL—it optimizes the application of existing knowledge but cannot supply missing domain knowledge. In contrast, SFT excels at injecting procedural knowledge through the direct imitation of expert reasoning traces. These approaches are thus complementary: SFT provides the *"content to know"*, while AT-RL dictates *"how to ground it correctly"* by reinforcing perceptual anchors. Integrating both is essential to resolve knowledge deficits and grounding deficiencies simultaneously. As shown in our analysis, AT-RL achieves **30.3%** accuracy on MathVision, specifically mitigating perception failures from the **22.7%** zero-shot baseline.

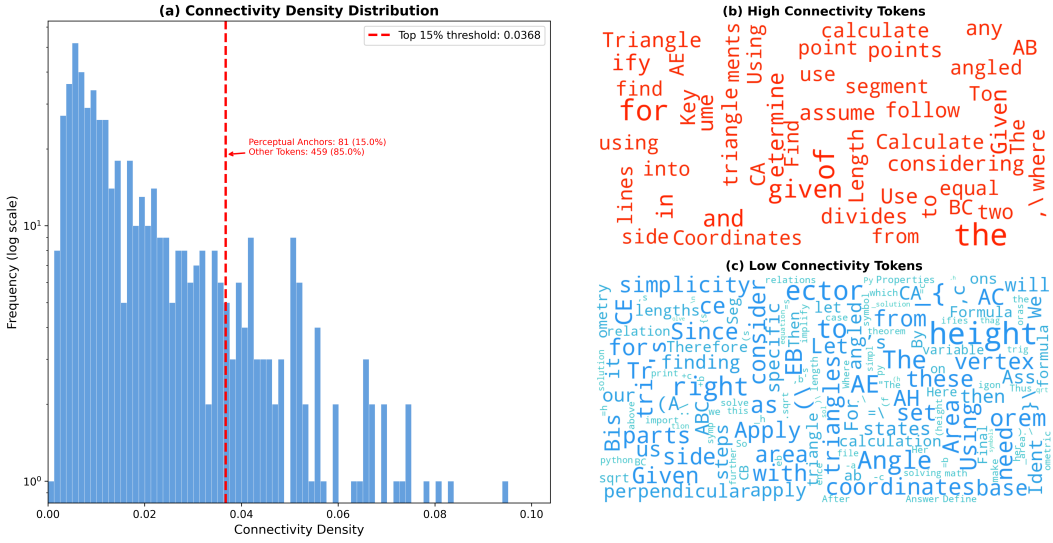

Figure 3: Cross-modal connectivity patterns. (a) Skewed distribution of connectivity density across tokens. (b) & (c) Word clouds of top 100 high vs. low connectivity tokens. Font size indicates average density.

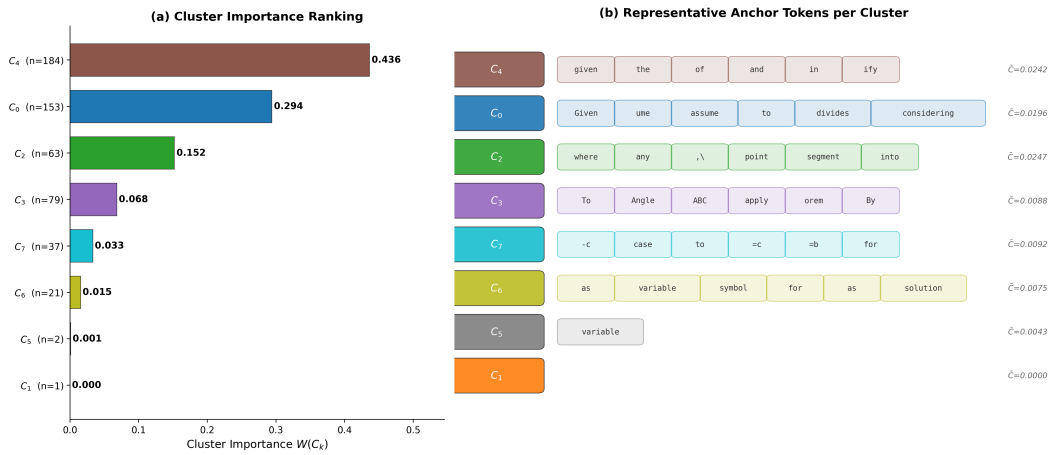

Figure 4: Analysis of perceptual anchor clusters in MathVision. (a) Cluster importance ranking. (b) Representative tokens for high-importance (G_4, G_0) vs. low-importance clusters.

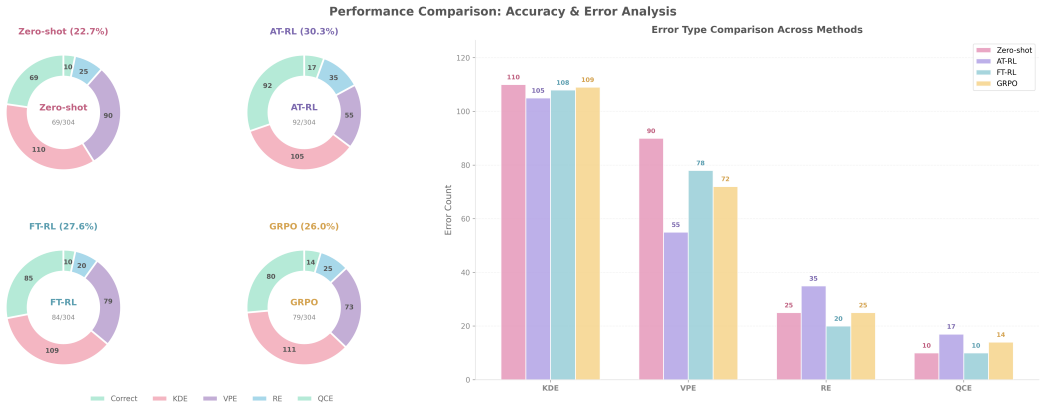

Figure 5: Performance and error breakdown on MathVision. (Left) Accuracy improvement: 22.7% (Zero-shot) $\rightarrow$ 30.3% (AT-RL). (Right) Absolute error counts by type, highlighting the significant reduction in perception errors (VPE).

Discussion 4: Credit Goes to Anchors — Whether Right or Wrong. AT-RL leverages the verifiable reward signal from RLVR to assign credit based on cross-modal connectivity, regardless of whether the final answer is correct or incorrect. When the model produces a correct answer, the positive reward reinforces high-connectivity tokens (perceptual anchors), strengthening their role in grounding reasoning on visual evidence. Conversely, when the answer is wrong, the negative reward is also concentrated on these same anchors, as they are most likely responsible for misinterpreting the visual input. This mechanism ensures that learning signals are consistently aligned with the fidelity of visual grounding, rather than being diluted across all tokens or biased by linguistic priors. In contrast, uniform credit assignment or SFT cannot distinguish the functional role of individual tokens in multimodal reasoning, leading to less efficient and less reliable updates.

B Implementation Details

This appendix contains engineering details, extended experiments, and formulae omitted from the main text for brevity. All components described here are necessary for full reproduction of **AT-RL** (**A**nchor-**T**oken **R**einforcement **L**earning).

Table 7: Sensitivity analysis of cluster refinement hyperparameters (τ_{cen}, α) on MathVerse and MathVista *testmini* using Qwen2.5-VL-7B-Instruct trained on Geometry-3K. The default setting is $\tau_{\text{cen}} = 0.75, \alpha = 0.6$.

	MathVerse		**MathVista**	
Configuration	Acc@1	Δ	Acc@1	Δ
Default ($\tau_{\text{cen}} = 0.75, \alpha = 0.6$)	**48.34**	**0.00**	**68.60**	**0.00**
$\tau_{\text{cen}} = 0.70, \alpha = 0.6$	47.92	-0.42	68.10	-0.50
$\tau_{\text{cen}} = 0.80, \alpha = 0.6$	47.85	-0.49	68.00	-0.60
$\tau_{\text{cen}} = 0.75, \alpha = 0.4$	47.70	-0.64	67.90	-0.70
$\tau_{\text{cen}} = 0.75, \alpha = 0.8$	48.05	-0.29	68.20	-0.40

Table 8: Sensitivity analysis of neighbourhood expansion hyperparameters (R, τ_{nb}, q) on MathVerse and MathVista *testmini* using Qwen2.5-VL-7B-Instruct trained on Geometry-3K. The default setting is $R = 4, \tau_{\text{nb}} = 0.65, q = 0.15$.

	MathVerse		**MathVista**	
Configuration	Acc@1	Δ	Acc@1	Δ
Default ($R = 4, \tau_{\text{nb}} = 0.65, q = 0.15$)	**48.34**	**0.00**	**68.60**	**0.00**
$R = 3, \tau_{\text{nb}} = 0.65, q = 0.15$	48.00	-0.34	68.20	-0.40
$R = 5, \tau_{\text{nb}} = 0.65, q = 0.15$	48.15	-0.19	68.40	-0.20
$R = 4, \tau_{\text{nb}} = 0.60, q = 0.15$	47.90	-0.44	68.10	-0.50
$R = 4, \tau_{\text{nb}} = 0.70, q = 0.15$	48.10	-0.24	68.30	-0.30
$R = 4, \tau_{\text{nb}} = 0.65, q = 0.10$	48.05	-0.29	68.25	-0.35
$R = 4, \tau_{\text{nb}} = 0.65, q = 0.20$	48.12	-0.22	68.35	-0.25

B.1 Bias Correction Details

The bias curve $b \in \mathbb{R}^{|\mathcal{V}|}$ is parameterised as:

$$\begin{aligned} b_j &= 1 + \lambda_{\text{exp}} e^{-p_j \gamma} + \lambda_{\cos} \cos(\pi p_j), \\ p_j &= \frac{j}{|\mathcal{V}|}, \quad j = 1, \ldots, |\mathcal{V}|, \end{aligned} \tag{12}$$

where λ_{exp} controls exponential term weight, γ decay rate, and $\lambda_{\cos}$ cosine amplitude. Mean-normalise b so $\frac{1}{|\mathcal{V}|} \sum_{j=1}^{|\mathcal{V}|} b_j = 1$, then divide $\bar{\mathbf{A}}$ element-wise to obtain $\bar{\mathbf{A}}^{\beta}$.

Default: $\lambda_{\text{exp}} = 0.15, \gamma = 4.0, \lambda_{\cos} = 0.05$.

B.2 Cluster Refinement

Given an initial METIS clustering $\{G_k\}$:

$$\mathbf{c}_k = \frac{1}{|G_k|} \sum_{t \in G_k} \bar{\mathbf{A}}_t^{\beta} \tag{13}$$

The credit weights of tokens with $\cos(\bar{\mathbf{A}}_t^{\beta}, \mathbf{c}_k) < \tau_{\text{cen}}$ are down-weighted:

$$\phi_t \leftarrow \begin{cases} \phi_t, & \cos(\bar{\mathbf{A}}_t^{\beta}, \mathbf{c}_k) \geq \tau_{\text{cen}} \\ \alpha \cdot \phi_t, & \text{otherwise} \end{cases} \tag{14}$$

Default: $\tau_{\text{cen}} = 0.75, \alpha = 0.6$.

B.3 Neighbourhood Expansion

Top-q percentile tokens (by ϕ_t) promote R nearest neighbours if:

$$\lambda_{\text{sim}} \cos(\bar{\mathbf{A}}_{t'}^{\beta}, \bar{\mathbf{A}}_t^{\beta}) + \lambda_{\text{imp}} \phi_{t'} > \tau_{\text{nb}} \tag{15}$$

Default: $q = 0.15$, $R = 4$, $\lambda_{\text{sim}} = 0.5$, $\lambda_{\text{imp}} = 0.5$, $\tau_{\text{nb}} = 0.65$.

B.4 Hyperparameters for AT-RL Optimisation

Batch size 512, Group size $G = 8$, group-wise mean/std advantage normalisation, PPO clip $\epsilon = 0.2$, KL coef $\beta = 0.02$, AdamW lr 1×10^{-6}, wd 0.01, 1 update/rollout.

B.5 Notation

$c_t^{(i)}$: token credit from $W(G_k)$; $A^{(i)}$: group-relative sequence advantage; $A_{\text{AT}}^{(i,t)}$: token-level advantage; $r_t(\theta)$: token-level probability ratio; τ_{sim}: graph edge threshold; ϕ_t: graph importance score.

B.6 Graph Construction Hyperparameters

$\tau_{\text{sim}} = 0.7$, $K = \max(2, \lfloor T/10 \rfloor)$.

C Applicability and Plug-and-Play Nature

C.1 KL-Independence and Flexibility

AT-RL is fundamentally a plug-and-play token-level advantage re-weighting mechanism that is completely agnostic to the choice of regularization strategy. While our default implementation includes a KL regulariser ($\beta = 0.02$) to stabilise policy updates, this is purely optional and not intrinsic to the AT-RL methodology.

The core innovation of AT-RL lies in its anchor-token credit assignment, which operates independently of any divergence control mechanism. This design philosophy enables seamless integration into diverse RL paradigms:

- **KL-Free Optimization**: Set $\beta = 0$ to completely remove KL regularization, as demonstrated in our ablation studies (Section C.2).
- **Alternative Regularizers**: Replace KL with entropy bonuses or trust-region constraints without modifying AT-RL's core logic.
- **Cross-Framework Compatibility**: Integrate into REINFORCE, actor-critic, DPO, or KL-free methods like DAPO by simply substituting the advantage term with $A_{\text{AT}}^{(i,t)}$.

Key Insight: AT-RL's token-level credit assignment $c_t^{(i)}$ derived from semantic clustering is orthogonal to the policy update rule. Any RL algorithm using per-token advantages can directly benefit from AT-RL's refined credit signal.

C.2 KL-Free AT-RL Experiments

To validate AT-RL's independence from KL regularization, we trained AT-RL(+GRPO) with $\beta = 0$:

Table 9: Acc@1(%) comparison between KL-regularised and KL-free AT-RL(+GRPO). Performance remains comparable, confirming that AT-RL's effectiveness stems from semantic credit assignment.

Method	$\text{GeoQA}_{\text{test}}$	MathVerse
AT-RL ($\beta = 0.02$)	45.41	40.68
AT-RL ($\beta = 0$)	45.27	40.55

The negligible performance difference ($< 0.2\%$) demonstrates that AT-RL's gains are primarily attributable to its anchor-grounded credit assignment mechanism.

C.3 Integration with KL-Free Methods

AT-RL can be directly integrated into recent KL-free optimization methods:

DAPO (Divergence-Agnostic Policy Optimization) To integrate AT-RL:

$$\mathcal{L}_{\text{DAPO+AT-RL}} = \mathbb{E}_t\left[\min\left(r_t(\theta)A_{\text{AT}}^{(i,t)}, \text{clip}(r_t(\theta), 1-\epsilon_t, 1+\epsilon_t)A_{\text{AT}}^{(i,t)}\right)\right] \tag{16}$$

where $A_{\text{AT}}^{(i,t)} = c_t^{(i)} \cdot A^{(i)}$ is the refined token-level advantage.

D AT-RL in Other RL Frameworks

D.1 REINFORCE

Replace the standard advantage $\hat{A}_t$ with AT-RL's token-level advantage:

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=1}^{T} A_{\text{AT}}^{(i,t)} \cdot \nabla_\theta \log \pi_\theta(a_t|s_t)\right] \tag{17}$$

D.2 Actor-Critic

Substitute the advantage term in the policy loss:

$$\mathcal{L}_{\text{policy}} = -\mathbb{E}\left[\sum_{t=1}^{T} A_{\text{AT}}^{(i,t)} \log \pi_\theta(a_t|s_t)\right] \tag{18}$$

D.3 DPO and Preference Optimization

Weight per-token log-likelihood differences by AT-RL credits:

$$\mathcal{L}_{\text{DPO+AT-RL}} = -\mathbb{E}\left[\log\sigma\left(\beta\sum_{t=1}^{T} c_t^{(i)} \cdot \left(\log\frac{\pi_\theta(y_w^t)}{\pi_{\text{ref}}(y_w^t)} - \log\frac{\pi_\theta(y_l^t)}{\pi_{\text{ref}}(y_l^t)}\right)\right)\right] \tag{19}$$

E Attention Extraction & Modality Labelling

We extract attention weights from the final $L = 4$ decoder layers, averaging across $H = 32$ heads. Visual tokens are assigned as v (vision) and text tokens as l (language).

F Environment and Complexity

Software Stack: PyTorch 2.2, CUDA 12.1, PyMetis 2023. **Hardware**: 8× NVIDIA A100-80GB GPUs.

Computational Breakdown per Training Iteration:

- Forward pass: 65.4%
- Backward pass: 33.5%
- AT-RL credit computation: 1.2%

The negligible overhead ($< 2\%$) confirms AT-RL's efficiency.

G Hyperparameter Search

We performed grid search over the following: $\tau_{\text{cen}} \in \{0.70, 0.75, 0.80\}$, $\alpha \in \{0.4, 0.6, 0.8\}$, $R \in \{3, 4, 5\}$, $\tau_{\text{nb}} \in \{0.60, 0.65, 0.70\}$, $q \in \{0.10, 0.15, 0.20\}$.

H Evaluation Protocol

All reported benchmark scores use the same task implementations and configuration of LMMs-Eval Zhang et al. (2024a) for every compared method. AT-RL changes training-time credit assignment only; it does not change prompts, decoding parameters, answer extraction, evaluator models, thresholds, or post-processing.

H.1 Task-specific answer evaluation

The evaluator follows each benchmark's official protocol. MathVision uses deterministic parsing and rule-based correctness. MathVista first extracts an answer using the benchmark parser and then applies normalized matching; an LLM fallback is used only when the official task implementation cannot deterministically extract an answer. MathVerse uses its official semantic-answer scoring protocol. We do not use a method-specific semantic judge.

H.2 Sampling configuration

Zero-shot evaluation uses three randomized prompt templates, temperature 0.1, and a maximum of 4096 generated tokens. For Mean@8, we run $N = 8$ independent repetitions:

$$\text{Mean@N} = \frac{\text{Number of passed problem instances}}{\text{Total evaluated problem instances}}. \tag{20}$$

H.3 Protocol audit

The same evaluator configuration is used for all paired comparisons. In a 200-example MathVista audit, the official answer-extraction procedure agreed with the manually checked extracted answer on more than 99.5% of examples. This audit does not eliminate benchmark-specific evaluator bias; it verifies that the reported method comparisons do not use different evaluators.

I Foundation model comparison

The complete comparison of Qwen2.5-VL-32B with larger and proprietary multimodal models is reported below.

Table 10: Comprehensive performance comparison of Qwen2.5-VL-32B with leading large multimodal models across mathematical, multi-discipline, and general VQA benchmarks.

Category	**Benchmark**	**Q2.5-32B** *Instruct*	**Q2.5-32B** *+Ours*	**Llama 4** *Scout*	**Claude 3.5** *Sonnet*	**Gemini 2.0** *Flash*	**OpenAI** *GPT-4o*	**Q2.5-72B** *Instruct*
Mathematical	MathVista_testmini	75.7	**80.2**	70.7	67.7	73.4	60.0	77.8
	MathVerse_testmini	49.3	**56.6**	–	47.8	54.6	41.2	57.2
	MathVision_test	38.2	**44.2**	–	33.5	41.3	30.6	43.1
Multi-Discipline	MMMU-Pro	48.5	**51.9**	52.2	51.5	51.7	51.9	51.6
	MMMU_Val	31.7	**36.5**	–	30.5	32.9	31.1	37.2
	EMMA_core	14.7	**19.4**	24.6	18.7	17.2	6.3	17.7
Real World	MEGA	48.9	**53.0**	31.8	52.5	54.7	52.7	49.6
Chart/VQA	ChartQAPro	43.4	**48.4**	–	55.8	53.7	46.9	47.2
	chatmimic	78.8	**79.4**	–	65.4	–	83.2	–
General VQA	TextVQA	84.6	**85.3**	62.5	75.2	78.1	80.5	84.3
	InfoVQA	–	**87.1**	60.1	72.8	75.6	77.0	76.9

J Additional Rebuttal Evidence

This appendix records the additional analyses performed during the review discussion. They are reported here so that the revised manuscript does not rely on claims that appeared only in the rebuttal.

Table 11: Inference-time patch deletion on MathVision. The result establishes causal importance of the selected image regions; it does not by itself establish token-level causal responsibility.

Masked patches	Accuracy	Δ vs. original
None (original)	30.3	0.0
Non-anchor-attended	29.1	-1.2
Random patches	26.1	-4.2
Anchor-attended	20.1	-10.2

Table 12: Alignment of visual attention with annotated task-relevant regions.

Token type	Top patch inside key region	Total attention inside key regions
Anchor	73.8%	57.4%
Non-anchor	15.0%	14.8%
Random baseline	17.5%	17.5%

J.1 Inference-time region intervention

On 200 MathVision samples, we aggregate the visual attention of the top-connectivity tokens and mask the top 20% of patches in that heatmap by replacing them with the image mean. The same mask ratio is used for non-anchor-attended patches and a random-patch control. The intervention is applied only at inference time and therefore does not change training.

J.2 Alignment with annotated key regions

We use Gemini-3-Pro to annotate boxes for key visual elements in 80 geometry samples and cross-check 30 samples with Claude-4.1-Opus. 91% of cross-check box pairs have IoU greater than 0.5. The annotations are used only for this diagnostic and are not used to train AT-RL.

J.3 Cross-architecture transfer

We apply the same training data and hyperparameters to two model families that use different visual encoders, language backbones, and visual-tokenization schemes.

These results support transfer across the tested families. They do not establish universality for architectures that do not expose cross-modal attention maps.

J.4 Three-seed variance

Table 14 reports mean $\pm$ standard deviation over three independent seeds for the Qwen2.5-VL-7B comparison. The full per-benchmark values are included to make the small number of seeds explicit.

The average gains over the corresponding base engines are +3.2, +3.6, and +3.4 points for GRPO, DAPO, and SAPO, respectively. Because only three seeds are available, we report standard deviations rather than making a stronger population-level significance claim.

J.5 Evaluation protocol and verifier audit

All methods are evaluated with the same version and task configuration of LMMs-Eval Zhang et al. (2024a). The framework has task-specific evaluators: deterministic parsing and rule-based matching are used where the benchmark defines them; an LLM is used only when the official task evaluator requires answer extraction or semantic equivalence. AT-RL does not modify prompts, evaluator models, thresholds, or post-processing. We therefore do not describe Qwen2.5-72B as a universal correctness judge. The previous draft's wording was too broad and has been removed.

For the reported benchmarks, MathVision uses rule-based correctness, MathVista uses answer extraction followed by normalized matching, and MathVerse follows its official semantic-answer pro-

Table 13: Transfer to non-Qwen architectures. Values are accuracy; Δ is AT-RL minus the corresponding GRPO result.

Model / method	MathVista	MathVision	MathVerse	Avg.
InternVL2.5-8B, GRPO	67.1	22.3	42.4	43.9
InternVL2.5-8B, + AT-RL	69.8	25.0	45.6	46.8
Delta	+2.7	+2.7	+3.2	+2.9
Qwen3-VL-8B, GRPO	78.8	55.7	64.3	66.3
Qwen3-VL-8B, + AT-RL	80.5	58.0	67.1	68.5
Delta	+1.7	+2.3	+2.8	+2.3

Table 14: Mean $\pm$ standard deviation over three seeds on Qwen2.5-VL-7B.

Method	GeoQA	MathVerse	MathVision	MathVista	WeMath	Avg.
GRPO	46.9±0.38	41.8±0.27	24.4±0.51	63.8±0.19	59.7±0.31	47.3±0.15
DAPO	47.5±0.33	42.6±0.31	25.4±0.43	64.3±0.25	60.4±0.22	48.0±0.17
SAPO	49.5±0.29	44.8±0.24	27.6±0.38	66.0±0.27	61.8±0.26	49.9±0.14
GRPO + AT-RL	49.2±0.35	46.0±0.28	28.0±0.44	66.7±0.21	62.7±0.23	50.5±0.16
DAPO + AT-RL	50.1±0.37	47.6±0.22	28.7±0.47	68.1±0.24	63.6±0.19	51.6±0.13
SAPO + AT-RL	52.8±0.31	48.3±0.26	30.3±0.42	68.6±0.18	66.2±0.28	53.3±0.15

tocol. The corresponding average AT-RL gains over the three base engines are +3.2, +3.1, and +4.2 points, respectively. A small audit of 200 MathVista examples reported over 99.5% answer-extraction agreement for the official parser. These checks do not rule out benchmark-specific evaluator bias; they show that the same evaluator configuration is applied to every paired method.

J.6 Qualitative failure analysis

Among 50 MathVision samples where GRPO fails and AT-RL succeeds, 27 corrected errors were categorized as visual-perception errors: visual-element identification (14/21 corrected), spatial relations (8/13), and visual hallucination (5/8). The remaining unresolved cases were mainly knowledge-deployment errors, such as selecting an inappropriate theorem after the relevant visual elements had been identified. This analysis supports the claim that AT-RL targets visual grounding; it does not claim that AT-RL supplies missing domain knowledge.

J.7 Scope and limitations

Cross-modal connectivity is an operational grounding signal, not a complete explanation of the model's reasoning. Patch deletion establishes the importance of selected image regions, while annotation agreement and the evaluator audit remain subject to model- and benchmark-specific bias. The experiments cover three model families, and the method requires access to cross-modal attention maps. The reported variance uses three seeds, and the RL procedure cannot add domain knowledge that is absent from the starting model. These limitations define the scope of the claims above.