PRPO: Perception-Reinforced Policy Optimization via Token-Level Dynamic Advantage Reshaping

Qiming Li[1,♠], **Tianlun Li**[1,†], **Xiaolong Cheng**[1],
Hangyu Li[2], **Ruiyan Gong**[1], **Kangning Niu**[1], **Kaitao Jiang**[1], **Mu Xu**[1]
[1]Amap CV Lab, Alibaba Group, [2]Peking University
qimingli1104@gmail.com

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has become an effective paradigm for improving the reasoning capability of Large Vision-Language Models (LVLMs). However, existing RLVR methods primarily rely on trajectory-level outcome rewards, which assign identical learning signals across all generated tokens. This coarse-grained credit assignment is fundamentally mismatched to multimodal reasoning, where only a sparse subset of tokens is causally grounded in visual evidence. Consequently, these pivotal perceptual tokens receive weak supervision and are often overwhelmed by language priors or reasoning-template tokens. To address this limitation, we propose **Perception-Reinforced Policy Optimization (PRPO)**, a token-level reinforcement learning framework that explicitly identifies and reinforces pivotal perceptual tokens within long-horizon multimodal reasoning trajectories. PRPO introduces **Robust Visual Dependency (RVD)**, a principled metric that identifies tokens whose predictions are both visually grounded and perturbation-stable, filtering out brittle or noisy visual tokens. Based on RVD, we further propose **Perceptual Advantage Reshaping (PAR)**, a token-level credit assignment technique that amplifies perceptually informative tokens while preserving stable gradients for non-perceptual tokens. Extensive experiments on seven multimodal reasoning benchmarks demonstrate that **PRPO** consistently outperforms strong LVLM baselines across both 3B and 7B model scales, achieving average gains of 23.3% and 21.1%, respectively. **PRPO** achieves state-of-the-art performance with improved training efficiency and stronger cross-task generalization. Our findings highlight the importance of fine-grained credit assignment for scalable multimodal reinforcement learning.

1 Introduction

Recent advances in reinforcement learning from verifiable rewards (RLVR) have significantly strengthened the reasoning capability of Large Language Models (LLMs), particularly when paired with online policy optimization algorithms such as Group Relative Policy Optimization (GRPO) [Shao et al., 2024, Guo et al., 2025, OpenAI, 2024, Team et al., 2025, Yang et al., 2025a, Anthropic, 2025, Zhang et al., 2025, Qu et al., 2025, Yan et al., 2025]. These gains have been most evident in text-focused tasks and benchmarks. Inspired by this success, researchers have increasingly explored how to adapt RLVR to Large Vision-Language Models (LVLMs). Current efforts can be broadly organized into three lines of work: methods that improve training data quality or composition [Li et al., 2025, Liang et al., 2025, Liu et al., 2025, Yao et al., 2025, Chen et al., 2025a, Meng et al., 2025, Huang et al., 2025b, Yang et al., 2025b]; methods centered on reward modeling, reward shaping, or

♠ Work done during an internship at Amap CV Lab, Alibaba Group.
† Corresponding author.

Preprint.

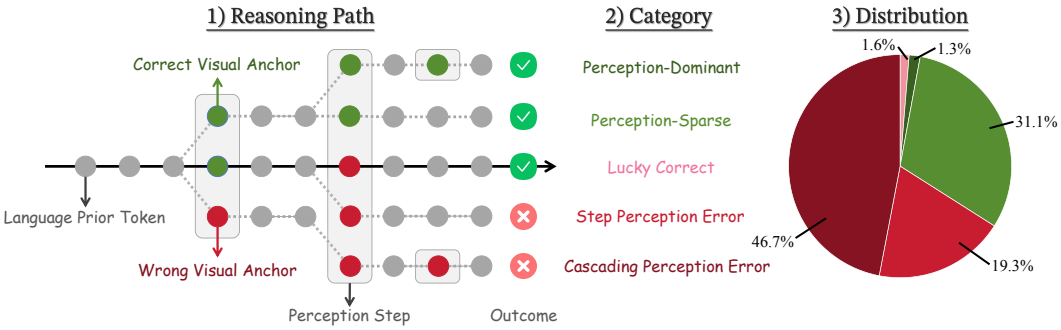


Figure 1: Error analysis of multimodal reasoning failures using Qwen2.5-VL-7B on the MathVerse benchmark. Analysis results indicate that LVLMs are limited by perceptual bottlenecks within reasoning trajectories rather than insufficient reasoning capabilities.

reward-system design [Shen et al., 2025, Xia et al., 2025, Wang et al., 2025b, Xiao et al., 2025, Yu et al., 2025a, Wan et al., 2025, He et al., 2025]; and methods that modify the underlying optimization procedure itself [Wang et al., 2025a, Zhao et al., 2025].

Despite these advances, existing multimodal RLVR methods largely inherit designs from language-only settings, which assign a single trajectory-level reward to the final response and broadcast the same optimization signal across all generated tokens. While this design is useful for purely textual reasoning, it becomes fundamentally mismatched in multimodal reasoning. To better characterize this mismatch, we conduct an error analysis on MathVerse [Zhang et al., 2024] using Qwen2.5-VL-7B. As shown in Figure 1, the majority of failures do not stem from insufficient reasoning capability, but from errors in visual perception occurring within long reasoning trajectories. Specifically, LVLMs often ground on incorrect visual regions, misread local evidence, or extract erroneous numerical cues at intermediate perception steps. Such behaviors either lead to isolated *Step Perception Errors* or accumulate into *Cascading Perception Errors*. In contrast, failures attributable purely to reasoning errors are relatively rare. Furthermore, we find that key perceptual tokens in multimodal reasoning (e.g., coordinates, numerical values, or geometric cues) are both sparse and highly influential, yet they receive the same supervision as non-perceptual linguistic tokens under standard RL objectives. We identify this mismatch as the ***Perceptual Credit Assignment Problem***, which contains two key issues: (i) sparse perceptual signals cannot be explicitly identified or reinforced from trajectory-level rewards, and (ii) uniform advantage assignment causes gradient dilution, where critical perceptual signals are overwhelmed by abundant reasoning tokens. As a result, existing RL methods tend to optimize surface-level reasoning fluency rather than improving visually grounded decision-making.

To address this problem, we propose **Perception-Reinforced Policy Optimization (PRPO)**, a dynamic token-level reinforcement learning framework that explicitly targets perceptual credit assignment in multimodal reasoning. **PRPO** contains two key components: **(i) Robust Visual Dependency (RVD)**, identifies tokens that are genuinely grounded in visual evidence and robust to perturbations, and **(ii) Perceptual Advantage Reshaping (PAR)**, which reshapes trajectory-level advantage into fine-grained token-level learning signals based on RVD, amplifying informative perceptual tokens while preserving stable optimization gradients on reasoning-template tokens.

Extensive experiments on seven multimodal reasoning benchmarks demonstrate that **PRPO** consistently improves performance across both 3B and 7B models, with average gains of 23.3% and 21.1%, respectively. Moreover, **PRPO** exhibits stronger performance under the same training budget and cross-task generalization, indicating that fine-grained perceptual reinforcement is more effective than uniform trajectory-level optimization in multimodal RLVR context.

Our contributions are three-fold:

- We identify ***Perceptual Credit Assignment*** as a key bottleneck in multimodal RLVR: most failures arise from sparse but critical perception errors rather than insufficient reasoning capability, while existing trajectory-level rewards fail to provide effective supervision to pivotal perceptual tokens.
- We propose **PRPO**, a token-level RL framework with **RVD** and **PAR** technique to identify and reinforce pivotal perceptual tokens while preserving stable optimization gradients.

- We achieve state-of-the-art (SOTA) results across seven multimodal reasoning benchmarks, demonstrating the importance of fine-grained credit assignment for multimodal reinforcement learning.

2 Related Work

Dominant Strategies in Multimodal RL. Current efforts to enhance multimodal RL largely focus on augmenting elements outside the core optimization algorithm. These include data-driven methods emphasizing the construction of visually-grounded datasets [Bai et al., 2025, Li et al., 2025, Liang et al., 2025] and CoT distillation [Chen et al., 2025b, Huang et al., 2025b, Meng et al., 2025], as well as reward-centric methodologies designing improved feedback mechanisms [Ma et al., 2025, Fan et al., 2025, Xia et al., 2025]. Recognizing the vision-language gap, recent works like Perception-Aware Policy Optimization (PAPO) [Wang et al., 2025d] explicitly account for modality differences to engineer perception-sensitive signals. However, even with such modality-aware designs or modality-agnostic algorithmic advances like DAPO [Yu et al., 2025b], these strategies still broadcast a uniform learning signal across the entire sequence. This indiscriminate credit assignment creates a severe bottleneck where critical, sparse visual signals are easily submerged by high-volume linguistic priors.

Pivotal Tokens Identification in LVLMs. To mitigate gradient dilution, it is crucial to isolate specific perceptual tokens from critical reasoning. In text-only domains, this involves identifying high-entropy "forking points" [Wang et al., 2025c] or low-confidence errors [Vassoyan et al., 2025]. For LVLMs, approaches like VPPO [Huang et al., 2025a] attempt to address this by utilizing KL Divergence-based visual dependency to reweight trajectory-level advantages, aiming to focus gradient updates on visually-grounded steps. However, relying on a single visual dependency metric risks amplifying noise-sensitive tokens, and trajectory-level reward reshaping is fundamentally insufficient to precisely reinforce visually important tokens. To tackle this, our PRPO framework introduces a precise dynamic token-level optimization paradigm. By jointly evaluating visual dependency and consistency, we systematically filter out stochastic noise to isolate genuinely pivotal perceptual tokens. We then dynamically reshape their learning advantages, explicitly amplifying critical perceptual signals without triggering gradient collapse or compromising foundational reasoning capabilities.

3 Preliminaries: Group-based Policy Optimization

In multimodal reinforcement learning from verifier feedback (RLVF), we consider a dataset $\mathcal{D}$ where each sample consists of a visual input I, a textual query q, and a ground-truth answer a. Let π_θ denote the policy parameters of a LVLM.

Group Relative Policy Optimization (GRPO) GRPO [Shao et al., 2024] serves as a value-model-free baseline that estimates advantages through intra-group relative performance. For a prompt (I, q), the model generates a group of G responses $\{o_i\}_{i=1}^{G}$, each receiving a binary outcome reward $R_i \in \{0, 1\}$ based on its equivalence to a. The advantage $\hat{A}_i$ is computed by normalizing rewards within the group:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_k\}_{k=1}^{G})}{\text{std}(\{R_k\}_{k=1}^{G}) + \eta}, \quad \text{where } R_i = \begin{cases} 1.0, & \text{if eq}(a, o_i) \\ 0.0, & \text{otherwise} \end{cases} \tag{1}$$

where η is a small constant for numerical stability. The objective is to maximize the clipped surrogate loss with an integrated KL penalty:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(r_t^i(\theta)\hat{A}_i, \text{clip}(r_t^i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right)\right)\right]$$

$$\text{with } r_t^i(\theta) = \frac{\pi_\theta(o_t^i|I, q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t^i|I, q, o_{<t})}. \tag{2}$$

Decoupled Clip and Dynamic Sampling (DAPO) DAPO [Yu et al., 2025b] refines the optimization landscape by introducing an asymmetric clipping mechanism and dynamic sampling strategies. It

utilizes a token-level objective with decoupled thresholds $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$ and global length normalization:

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_t^i(\theta)\hat{A}_i,\ \text{clip}\big(r_t^i(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\big)\hat{A}_i\right)\right] \tag{3}$$

4 PRPO: Perception-Reinforced Policy Optimization

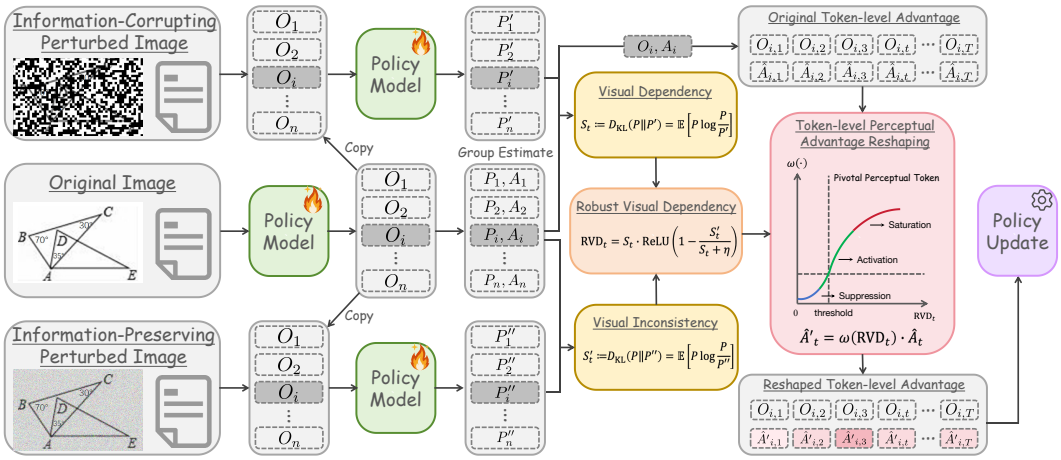


Figure 2: Overview of our proposed PRPO framework. PRPO consists of two core components: **(i) Robust Visual Dependency (RVD)**, which identifies *pivotal perceptual tokens* through fine-grained trajectory analysis, and **(ii) Perceptual Advantage Reshaping (PAR)** algorithm, which applies a synaptic S-shaped curve to reshape uniform trajectory-level advantages, dynamically amplifying pivotal perceptual tokens while preserving stable optimization gradients for reasoning tokens.

4.1 Robust Visual Dependency (RVD)

In this section, we detail the methodology for identifying *pivotal perceptual tokens* within long multimodal reasoning trajectories. Following established paradigms [Huang et al., 2025a], we adopt Kullback-Leibler (KL) divergence as a principled heuristic metric to quantify the sensitivity of individual tokens to visual evidence, serving as the foundation for our robust dependency analysis.

Definition 4.1 (Visual Dependency). *Let $\mathcal{P}_{str}$ be a strong perturbation operator that maps I to a state $\tilde{I}_{str}$ where the visual information is severely corrupted (e.g., via a stochastic black mask). We define the Visual Dependency S_t to formally quantify the policy's dependency on visual information for a specific output token o_t:*

$$S_t \coloneqq D_{\text{KL}}\big(\pi_\theta(\cdot \mid I, q, o_{<t}) \,\|\, \pi_\theta(\cdot \mid \mathcal{P}_{\text{str}}(I), q, o_{<t})\big) \tag{4}$$

S_t measures the absolute magnitude of the visual signal required for prediction. A high S_t indicates that the visual input I serves as a necessary perception anchor for the generation of token o_t, distinguishing visual grounded tokens from pure language priors.

Definition 4.2 (Visual Inconsistency). *Let $\mathcal{P}_{wk}$ be a weak perturbation operator that introduces an infinitesimal stochastic jitter $\boldsymbol{\delta} \sim \mathcal{N}(0, \sigma^2 \cdot I)$ to the pixel space. Crucially, $\mathcal{P}_{wk}$ ensures that the visual semantics remain strictly invariant. We define the Visual Inconsistency S'_t to measure the model's sensitivity to numerical noise for a specific output token o_t:*

$$S'_t \coloneqq D_{\text{KL}}\big(\pi_\theta(\cdot \mid I, q, o_{<t}) \,\|\, \pi_\theta(\cdot \mid \mathcal{P}_{\text{wk}}(I), q, o_{<t})\big) \tag{5}$$

S'_t probes the local curvature of the policy distribution on the perceptual manifold. It functions as a stability probe to identify numerical instability rather than semantic shift. A robust *pivotal perceptual tokens* must exhibit high S_t (strong *Visual Dependency*) but low S'_t (weak *Visual Inconsistency*).

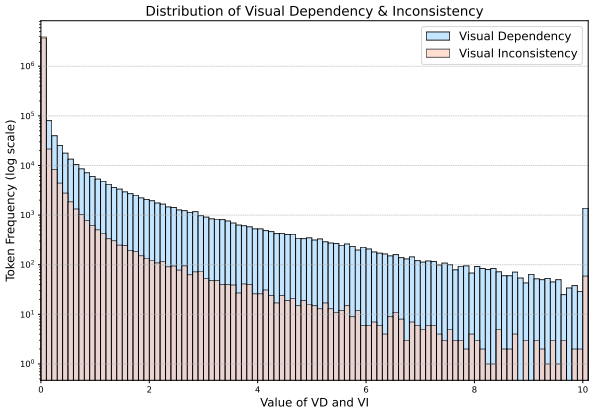


Figure 3: Distribution of Visual Dependency (VD) and Visual Inconsistency (VI) on Qwen2.5-VL-7B.

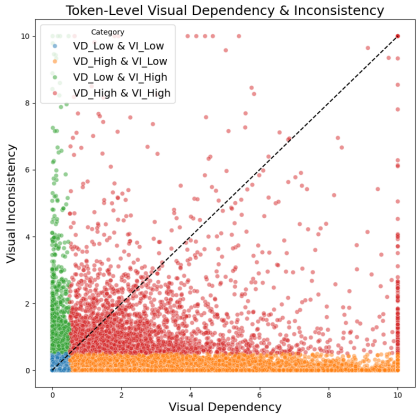


Figure 4: Token-Level distribution of VD and VI on Qwen2.5-VL-7B.

To provide a principled foundation for PRPO, we conduct a fine-grained diagnostic analysis on the MathVerse benchmark using Qwen2.5-VL-7B with a rollout factor of 8. By computing S_t and S'_t for each tokens, we uncover two critical findings regarding the distribution of visual signals:

Finding 4.1: Visual Signal Sparsity. Our analysis reveals that perceptual tokens are extremely sparse within long-horizon reasoning trajectories. As illustrated in Figure 3, both S_t and S'_t exhibit a dramatic **long-tail characteristic**. The vast majority of generated tokens reside in the low-value region ($S_t, S'_t < 0.1$), suggesting that the policy is predominantly driven by language priors rather than visual evidence. Traditional RL methods (e.g., GRPO and DAPO) that broadcast advantages uniformly across the trajectory suffer from **gradient dilution** in multi-modal context, where rare but critical visual signals are submerged by linguistic background noise.

Finding 4.2: The Coupled Nature of Visual Dependency & Inconsistency. Visual Dependency (S_t) and Inconsistency (S'_t) are inherently coupled rather than trivially separable. By projecting tokens across both dimensions (Figure 4), we can roughly partition them into three distinct regimes:

- **Linguistic Tokens (low S_t).** The vast majority of tokens fall into this region (the blue cluster). Unaffected by either perturbation, they are governed entirely by linguistic priors.
- **Brittle Anchors (high S_t, high S'_t).** A substantial subset of tokens exhibits strong visual dependency but equally high sensitivity to numerically unstable pixel-level features (the red cluster). Reinforcing them solely based on S_t would amplify noise rather than informative perception.
- **Pivotal Perceptual Tokens (high S_t, low S'_t).** The true *pivotal perceptual tokens* form a narrow and highly structured subset (the orange cluster). These tokens exhibit strong semantic grounding while remaining stable under local perturbations, indicating that their predictions are driven by informative visual evidence. We refer to these tokens as *pivotal perceptual tokens*, as they represent the causal interface where visual evidence is incorporated into the reasoning trajectory.

This decomposition exposes a fundamental limitation of pivotal token selection strategies, such as Jensen-Shannon Divergence, single KL Divergence [Huang et al., 2025a] and hidden state similarity [Li et al., 2026], which fail to distinguish **semantically grounded perception** from **numerically brittle responses**. A dual-metric synergistic approach ($S_t \uparrow, S'_t \downarrow$) is necessary to filter out stable *pivotal perceptual tokens* that drive visually-grounded reasoning. Further performance analysis of these metrics is provided in the Appendix B.2.

Definition 4.3 (Robust Visual Dependency). *Let S_t and S'_t be the Visual Dependency and Visual Inconsistency at step of token t, respectively. The hyperparameter η ensures numerical stability and serves as a baseline sensitivity threshold. We define the Robust Visual Dependency (RVD_t) to precisely quantify the real contribution of visual perception to the generation of token o_t:*

$$\mathrm{RVD}_t = S_t \cdot \mathrm{ReLU}\left(1 - \frac{S'_t}{S_t + \eta}\right) \tag{6}$$

The RVD index acts as a non-linear information filter. By utilizing the ReLU-gated ratio, it effectively suppresses tokens in the **stochastically brittle** region (where $S'_t \geq S_t$) while preserving the original magnitude of **Golden Anchors** (where $S'_t \ll S_t$). This ensures that the subsequent reinforcement signal is concentrated on tokens that are both semantically grounded and numerically robust.

4.2 Token-Level Dynamic Perceptual Advantage Reshaping (PAR)

While RVD provides a fine-grained measure of perceptual relevance, directly using it for credit assignment leads to suboptimal optimization behavior due to its highly skewed distribution:

(1) Long-tailed distribution. RVD exhibits a heavy-tailed structure, where the majority of tokens concentrate near zero, while only a small subset of perceptually grounded tokens take large values. Directly using such signals results in inefficient credit allocation.

(2) Gradient collapse. A naive transformation (e.g., linear or monotonic scaling) causes most token-level weights to shrink toward zero, effectively collapsing gradients onto a sparse subset of visual tokens. This leads to a biased optimization process that over-focuses on perceptual tokens while neglecting linguistic reasoning, ultimately degrading general reasoning performance.

To address these issues, we propose **Perceptual Advantage Reshaping (PAR)**, a differentiable nonlinear mapping that transforms raw perceptual signals into bounded token-level weights. The resulting *S-shaped synaptic response curve* suppresses low-signal tokens, amplifies informative perceptual tokens beyond a threshold, and enforces smooth saturation, thereby preventing gradient explosion while preserving stable optimization.

Synaptic S-shaped Reshaping. Formally, let $x = \mathrm{RVD}_t \in \mathbb{R}^+$ denote the raw perceptual stimulus. To achieve a strictly bounded and mathematically smooth transformation, we formulate the reshaped advantage $\mathcal{T}(x)$ via a generalized logistic function:

$$\mathcal{T}(x) = m + \frac{M - m}{1 + \exp\left(-k \cdot (x - \mu)\right)} \tag{7}$$

where the synaptic activation is governed by four highly interpretable parameters: the activation threshold $\mu > 0$, the upper saturation bound M and the lower suppression baseline m ($M > m \geq 0$), and a scaling factor $k > 0$ that dictates the sharpness of the activation.

Token-level Dynamic Perceptual Advantage Reshaping (PAR). Standard trajectory-level reinforcement learning suffers from coarse credit assignment, uniformly broadcasting a scalar reward to all tokens regardless of their perceptual contributions. To resolve this, we leverage our synaptic response function $\mathcal{T}(\cdot)$ to establish **PAR**:

$$\hat{A}_t^{i,\mathrm{PAR}} = \hat{A}^i \cdot \mathcal{T}(\mathrm{RVD}_t^i) \tag{8}$$

This mechanism organically redistributes the learning signals. Tokens exhibiting strong perceptual relevance (post-threshold) receive amplified advantages, directing the optimization focus toward semantically pivotal regions. Conversely, perceptual noise is softly discounted. Critically, the strictly positive lower bound m in $\mathcal{T}(\cdot)$ guarantees a non-zero gradient even for suppressed tokens, preserving dense learning signals and preventing optimization dead zones across the sequence.

Training Objectives. We formulate our final policy objective by integrating **PAR** into a clipped surrogate framework for robust optimization:

$$\mathcal{J}_{\mathrm{PRPO}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G}|o^i|}\sum_{i=1}^{G}\sum_{t=1}^{|o^i|}\min\left(r_t^i(\theta)\hat{A}_t^{i,\mathrm{PAR}},\ \mathrm{clip}\big(r_t^i(\theta), 1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}}\big)\hat{A}_t^{i,\mathrm{PAR}}\right)\right] \tag{9}$$

By bounding dynamically reshaped advantages within an asymmetric trust region, PRPO achieves precise multimodal alignment without overfitting to visual noise or compromising language priors.

5 Experiments

Table 1: Main results (Rollout $= 8$) on various benchmarks. The best results are **bolded**, and the second-best results are underlined. Comparison is performed separately within 3B and 7B groups.

Model	Geo3k	MathVista	We-Math	MMK12	MathVerse	Logic	MMMU-Pro	MathVerse$_V$	AVG
3B Models									
Qwen2.5-VL-3B	18.8	46.4	29.8	37.1	33.8	30.8	19.3	30.1	30.7
+ GRPO	28.7	59.3	58.9	57.2	55.3	38.1	25.7	52.3	46.9
+ DAPO	31.2	60.9	60.0	66.8	56.3	40.7	28.4	53.1	49.7
+ PAPO-G	31.1	59.7	59.9	56.3	57.5	39.5	26.4	53.6	48.0
+ PAPO-D	35.7	62.5	62.7	64.1	60.5	41.7	28.8	57.7	51.7
+ VPPO	35.8	63.6	61.4	65.3	61.3	42.0	28.8	58.1	52.1
+ PRPO(Ours)	**37.3**	**65.3**	**63.9**	**66.9**	**62.7**	**44.9**	**30.3**	**60.7**	**54.0**
7B Models									
MM-Eureka-7B	40.3	71.9	66.1	67.7	67.2	46.3	30.3	62.4	56.5
VL-Rethinker-7B	40.7	70.6	69.3	68.3	68.8	46.5	37.1	65.0	58.6
NoisyRollout-7B	**52.0**	68.6	69.6	51.0	67.8	47.5	34.5	63.8	56.8
R1-ShareVL-7B	41.2	69.3	68.7	70.9	68.2	45.8	35.1	64.3	57.9
Qwen2.5-VL-7B	36.7	63.6	46.7	43.3	39.5	42.0	25.2	34.3	41.4
+ GRPO	40.2	65.5	68.1	72.3	66.5	45.6	35.2	61.7	56.9
+ DAPO	35.9	61.9	58.5	75.9	55.6	37.1	29.0	51.0	50.6
+ PAPO-G	40.3	68.1	66.8	71.8	68.3	45.8	35.5	64.1	57.6
+ PAPO-D	44.1	67.5	68.3	80.6	68.6	46.7	36.3	64.9	59.6
+ VPPO	44.4	70.5	68.9	80.1	71.0	47.5	38.8	65.5	60.8
+ PRPO(Ours)	46.9	**72.2**	**70.8**	**82.8**	**72.2**	**49.7**	**38.8**	**66.3**	**62.5**

5.1 Experimental Setup

Baseline Models, Data and Methods. To ensure a rigorous and equitable comparison with prior SOTA approaches, we deliberately align our experimental setup—including backbone architectures and training corpora-with established conventions in recent literature. Specifically, we instantiate our PRPO framework utilizing the Qwen2.5-VL series (both 3B and 7B variants) as the base policy models, and conduct training exclusively on the `ViRL39K` [Wang et al., 2025a] dataset. Furthermore, we benchmark PRPO against nine prominent training-based methods, encompassing both foundational reinforcement learning paradigms and recent multimodal-specific advancements. These baselines include GRPO [Shao et al., 2024], DAPO [Yu et al., 2025b], PAPO [Wang et al., 2025d], VPPO [Huang et al., 2025a], MM-Eureka-7B [Meng et al., 2025], VL-Rethinker-7B [Wang et al., 2025a], NoisyRollout-7B [Liu et al., 2025], and R1-ShareVL-7B [Yao et al., 2025].

Evaluation Benchmarks. For a comprehensive evaluation of model capabilities, we assess performance across seven widely adopted benchmarks spanning complex mathematical reasoning and general multimodal understanding: Geo3K [Lu et al., 2021], MathVista [Lu et al., 2023], We-Math [Qiao et al., 2024], MMK12 [Meng et al., 2025], MathVerse [Zhang et al., 2024], LogicVista [Xiao et al., 2024], MMMU-Pro [Yue et al., 2024]. MathVerse$_V$ denotes the vision-centric subset of MathVerse.

Training Details. To ensure a rigorous and equitable comparison with baseline training methodologies, we align our experimental protocol with established settings of prior works. Specifically, both the 3B and 7B models are trained on the `ViRL39K` [Wang et al., 2025a] dataset for 2 epochs, employing a learning rate of 1e-6 and a rollout batch size of 384. The maximum response length is constrained to 2048 tokens. Comprehensive hyperparameter settings are detailed in Appendix A.3.

5.2 Experimental Results

As presented in Table 1, our proposed PRPO establishes SOTA performance across both 3B and 7B models, demonstrating comprehensive superiority over existing multimodal RL algorithms.

SOTA Performance and Consistency. PRPO achieves the best average performance on both the 3B and 7B models. Compared to the baseline Qwen2.5-VL Series, our models deliver substantial improvements of +23.3% (from 30.7% to 54.0%) and +21.1% (from 41.4% to 62.5%) at the 3B and 7B scales, respectively. Compared to the strongest baseline method VPPO, our models exhibit improvements of +1.9% and +1.7%. Furthermore, we achieve consistent performance gains across multiple tasks, demonstrating that PRPO significantly outperforms existing training methods.

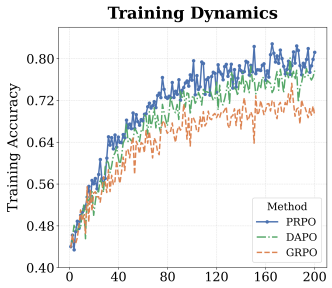


Figure 5: Training dynamics.

Dynamic Token-level Credit Assignment. A key distinction between PRPO and high-performing baselines like PAPO and VPPO lies in the granularity of credit assignment. While existing methods primarily rely on coarse **trajectory-level** reward signals, PRPO introduces a **dynamic token-level** advantage reshaping framework. This empirical gap validates that our proposed **RVD** and **PAR** techniques not only accurately identify *perceptual pivotal tokens* on line, but also effectively amplifying those tokens while suppressing irrelevant noise of language priors.

Superior Training Efficiency. As shown in Figure 5, PRPO exhibits significantly higher training efficiency than the GRPO and DAPO baselines. Under identical training step constraints, PRPO achieves markedly higher performance gains and exhibits faster convergence. This evidence demonstrates that the synaptic-inspired response curve provides high-quality, informative training signals, making PRPO a more efficient and scalable objective for the complex optimization landscape of LVLMs. See Appendix C for additional training efficiency analysis.

5.3 Ablation Study

Table 2: Ablation study of core components on Qwen2.5-VL-7B (Rollout $= 8$). The best results are **bolded**, and the second-best results are underlined.

Model	Geo3k	MathVista	We-Math	MMK12	MathVerse	Logic	MMMU-Pro	MathVerse_V	AVG
Qwen2.5-VL-7B	36.7	63.6	46.7	43.3	39.5	42.0	25.2	34.3	41.4
+ DAPO	35.9	61.9	58.5	75.9	55.6	37.1	29.0	51.0	50.6
+ PRPO w/o VI	44.8	71.0	69.2	80.5	71.3	48.1	38.6	65.8	61.2
+ PRPO w/o PAR	43.9	69.9	68.4	79.5	70.2	46.9	38.1	65.0	60.2
+ PRPO (Ours)	**46.9**	**72.2**	**70.8**	**82.8**	**72.2**	**49.7**	**38.8**	**66.3**	**62.5**

Ablation Study of Core Components. Table 2 demonstrates that both the Visual Inconsistency (VI) penalty and Perceptual Advantage Reshaping (PAR) are essential for multimodal RL. Removing the VI module (w/o VI) exposes the policy to single Visual Dependencies (VD), which increases hallucination in long reasoning chains (AVG drops to 61.2%). More importantly, removing the PAR mechanism (w/o PAR) causes a severe performance degradation to 60.2%. Without the bounded non-linear gating, large gradients from perceptual tokens dominate the linguistic priors, leading to representation collapse. Notably, even these ablated variants significantly outperform the trajectory-level DAPO baseline (50.6%), empirically demonstrating that fine-grained token-level routing is a fundamentally superior paradigm for multimodal RL.

Ablation Study of PAR Curve Parameters. We further investigate the sensitivity of synaptic gating function which is mainly shaped by (M, m, k) to ensure mathematical stability in optimization landscape. As shown in Table 3, we can summarize the specific effects of these parameters as follows:

- **Upper Saturation Bound (M):** A conservative bound ($M = 2.0$) inadequately amplifies critical tokens (AVG: 57.9%). Conversely, an excessive bound ($M = 6.0$) triggers reward dilution and high-variance updates. $M = 4.0$ optimally balances signal amplification and gradient regularization.
- **Lower Suppression Baseline (m):** This soft lower bound acts as a critical inductive bias against optimization dead zones. Over-suppressing it ($m = 0.05$) starves the gradient flow for subtle visual contexts, whereas elevating it ($m = 0.15$) paralyzes the noise-filtering capacity.

Table 3: Ablation study of key hyperparameters in Perceptual Advantage Reshaping (PAR) on Qwen2.5-VL-7B (Rollout $= 8$). The best results are **bolded**.

M	m	k	Geo3k	MathVista	We-Math	MMK12	MathVerse	Logic	MMMU-Pro	AVG
Baseline			36.7	63.6	46.7	43.3	39.5	42.0	25.2	42.4
Ablation on Upper Saturation Bound (M)										
2.0	0.10	2.0	43.2	68.5	65.2	79.5	67.1	46.1	35.8	57.9
6.0	0.10	2.0	44.5	69.8	67.5	78.0	68.5	47.2	36.5	58.9
Ablation on Lower Suppression Baseline (m)										
4.0	0.05	2.0	45.1	70.2	68.1	80.5	70.1	48.9	37.2	60.0
4.0	0.15	2.0	44.8	69.5	66.8	77.2	68.9	45.5	36.4	58.4
Ablation on Activation Sharpness (k)										
4.0	0.10	1.5	43.9	68.8	67.1	78.4	69.2	46.8	36.1	58.6
4.0	0.10	2.5	45.5	71.0	69.2	81.1	70.8	47.5	37.8	60.4
4.0	0.10	2.0	**46.9**	**72.2**	**70.8**	**82.8**	**72.2**	**49.7**	**38.8**	**61.9**

Table 4: Ablation study on the activation threshold (μ) percentile of RVD on Qwen2.5-VL-7B (Rollout $= 8$). The best results are **bolded**.

Threshold (μ)	Geo3k	MathVista	We-Math	MMK12	MathVerse	Logic	MMMU-Pro	AVG
Baseline	36.7	63.6	46.7	43.3	39.5	42.0	25.2	42.4
20th Percentile	46.1	71.5	69.5	81.6	71.4	48.8	38.0	61.0
25th Percentile	**46.9**	**72.2**	**70.8**	**82.8**	**72.2**	**49.7**	**38.8**	**61.9**
30th Percentile	45.2	70.4	68.3	80.2	70.1	47.9	37.3	59.9
40th Percentile	39.2	65.8	52.4	55.1	45.6	43.5	28.6	47.2

- **Activation Sharpness (k):** Modulating the phase transition of the gating mechanism, a dampened curve ($k = 1.5$) homogenizes the advantage distribution, failing to isolate critical tokens. In contrast, an aggressive step-like transition ($k = 2.5$) destabilizes optimization near the threshold. Setting $k = 2.0$ guarantees a smooth, robust landscape for optimal credit assignment.

Ablation Study of the Activation Threshold (μ). As shown in Table 4, we observe that setting the threshold μ to **the 25th percentile** of RVD yields the best performance. This indicates that reinforcing only the **Top-25%** of *pivotal perceptual tokens* is sufficient to achieve effective perceptual optimization, while maintaining a lower gradient level for the remaining 75% of tokens effectively filters linguistic noise and ensures the steady evolution of foundational reasoning capabilities.

5.4 Case Study: More Accurate Reasoning with Fewer Tokens

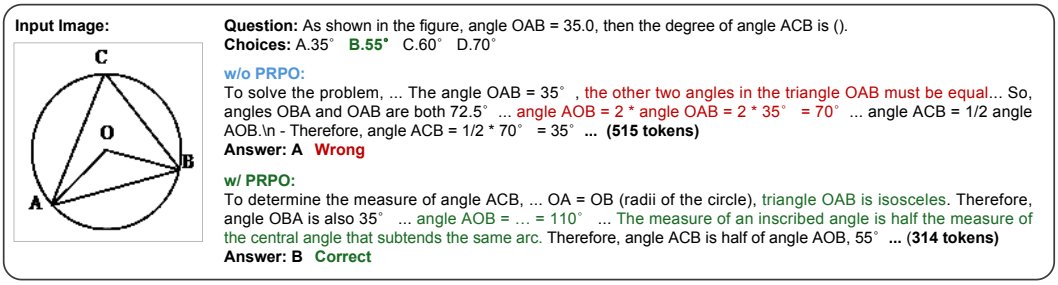


Figure 6: Comparison of the generation results between PRPO the baseline using Qwen2.5-VL-7B.

As shown in Figure 6, PRPO successfully corrects the perception errors of the baseline. Consequently, the entire reasoning trajectory becomes much more concise, significantly reducing required tokens.

6 Conclusion

We propose PRPO, a token-level reinforcement learning framework that addresses the Perceptual Credit Assignment Problem in multimodal reasoning. By introducing Robust Visual Dependency

(RVD) and Perceptual Advantage Reshaping (PAR), we move beyond coarse trajectory-level rewards to precisely reinforce pivotal perceptual tokens while filtering out numerical noise and linguistic priors. Our framework achieves SOTA performance across seven benchmarks, with average gains of 23.3% and 21.1% at 3B and 7B scales, respectively. These results demonstrate that fine-grained token-level credit assignment is essential for aligning multimodal reasoning systems at scale.

References

Anthropic. Claude sonnet 4, 2025. URL `https://www.anthropic.com/claude/sonnet`.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025.

Liang Chen, Hongcheng Gao, Tianyu Liu, Zhiqi Huang, Flood Sung, Xinyu Zhou, Yuxin Wu, and Baobao Chang. G1: Bootstrapping perception and reasoning abilities of vision-language model via reinforcement learning. *arXiv preprint arXiv:2505.13426*, 2025a.

Liang Chen, Lei Li, Haozhe Zhao, and Yifan Song. Vinci. r1-v: Reinforcing super generalization ability in vision-language models with less than $3, 2025b.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025.

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

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*, 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*, 2025a.

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

Shenshen Li, Kaiyuan Deng, Lei Wang, Hao Yang, Chong Peng, Peng Yan, Fumin Shen, Heng Tao Shen, and Xing Xu. Truth in the few: High-value data selection for efficient multi-modal reasoning. *arXiv preprint arXiv:2506.04755*, 2025.

Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, Zhaojie Liu, Qibin Hou, and Ming-Ming Cheng. Rethinking token-level policy optimization for multimodal chain-of-thought. *arXiv preprint arXiv:2603.22847*, 2026.

Yiqing Liang, Jielin Qiu, Wenhao Ding, Zuxin Liu, James Tompkin, Mengdi Xu, Mengzhou Xia, Zhengzhong Tu, Laixi Shi, and Jiacheng Zhu. Modomodo: Multi-domain data mixtures for multimodal llm reinforcement learning. *arXiv preprint arXiv:2505.24871*, 2025.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025.

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. *arXiv preprint arXiv:2105.04165*, 2021.

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. *arXiv preprint arXiv:2310.02255*, 2023.

Yan Ma, Linge Du, Xuyang Shen, Shaoxiang Chen, Pengfei Li, Qibing Ren, Lizhuang Ma, Yuchao Dai, Pengfei Liu, and Junjie Yan. One rl to see them all: Visual triple unified reinforcement learning. *arXiv preprint arXiv:2505.18129*, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

OpenAI. Learning to reason with llms, 2024. URL `https://openai.com/index/learning-to-reason-with-llms/`.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*, 2025.

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

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

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

Jean Vassoyan, Nathanaël Beau, and Roman Plaud. Ignore the kl penalty! boosting exploration on critical tokens to enhance rl fine-tuning. *arXiv preprint arXiv:2502.06533*, 2025.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*, 2025.

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. *arXiv preprint arXiv:2504.08837*, 2025a.

Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*, 2025b.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 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.

Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Advancing multimodal reasoning capabilities of multimodal large language models via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025b.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*, 2025.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025a.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025b.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

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

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.

A Experimental Details

A.1 Benchmarks and Evaluation Metrics

To provide a comprehensive evaluation and comparison of different LVLM RL methods, we align with prior literature and select the following benchmarks across diverse difficulty levels and domains:

(1) Geometric and Mathematical Reasoning. The primary focus of our experiments lies in assessing the model's ability to navigate various mathematical and geometric domains.

- **Geo3k** [Lu et al., 2021]: This benchmark specializes in geometry-centric problems at the high school level. Its rigorous use of formal language annotations allows for a precise evaluation of how well models can execute symbolic reasoning within a visual context.
- **MathVista**: Serving as a comprehensive evaluation of visually situated mathematics, this benchmark integrates a broad spectrum of visual contexts—ranging from statistical charts and functional plots to abstract geometry and real-world scenes. It systematically challenges models to synthesize complex visual evidence with deep quantitative and logical deduction, providing a robust measure of generalized mathematical capability across highly diverse visual modalities.
- **We-Math** [Qiao et al., 2024]: Focused on a human-centric diagnostic approach, We-Math breaks down complex problems into a hierarchy of 67 specific knowledge points. This granularity allows us to identify whether a failure stems from a lack of specific knowledge or a breakdown in the reasoning chain.
- **MMK12** [Meng et al., 2025]: This benchmark provides a standardized assessment of foundational math skills derived from K-12 curricula. It ensures that the model possesses the prerequisite quantitative literacy required for more specialized tasks.
- **MathVerse** [Zhang et al., 2024]: This benchmark probes the visual dependency of Large Vision-Language Models. By offering six variants of the same problem with differing ratios of visual-to-textual information, it uncovers whether a model truly "perceives" a diagram or simply exploits linguistic redundancies.

(2) Logical and Cognitive Foundations. Beyond pure mathematics, we evaluate the model's underlying cognitive architecture through targeted logical tasks.

- **LogicVista** [Xiao et al., 2024]: Designed to measure general-purpose logic, this benchmark transcends the mathematical domain by testing five core reasoning pillars: induction, deduction, and spatial, mechanical, and numerical reasoning. It offers a diverse array of visual formats to assess fundamental intellectual agility.

(3) Advanced Multimodal Integration. To measure the upper limits of university-level understanding where visual and textual data must be processed as a single entity, we utilize a specialized "pro" suite.

- **MMMU-Pro** [Yue et al., 2024]: As a more rigorous iteration of the MMMU dataset, this version is specifically engineered to defeat text-only shortcuts. By increasing the number of distractors and incorporating vision-exclusive formats, it demands that the model perform high-level academic reasoning that is fundamentally impossible without deep multimodal fusion.

A.2 Baseline Methods

Recently, researchers have optimized reinforcement learning for LVLMs from various perspectives. Specifically, we compare PRPO with the following closely related baseline methods:

- **PAPO** [Wang et al., 2025d] introduces a policy gradient algorithm that encourages reasoning steps to be internally grounded in visual inputs without relying on external annotations or reward models. However, its optimization objective is applied uniformly across all generated tokens, neglecting the inherently sparse nature of critical visual signals.
- **VPPO** [Huang et al., 2025a] attempts to address the uniform signal bottleneck through a two-tiered strategy, employing trajectory-level reward shaping combined with focusing policy updates on a sparse set of perception tokens.

Differences from our work: While both baselines seek to enhance visually-grounded reasoning, they exhibit key limitations that PRPO explicitly resolves. PAPO's uniform token treatment inevitably leads to gradient dilution, where critical visual signals are overwhelmed by abundant language priors. VPPO mitigates this by focusing on sparse tokens, but its reliance on coarse trajectory-level reward shaping and single-metric token selection risks amplifying numerically brittle visual noise.

In contrast, our PRPO framework operates through a highly precise, dynamic *token-level* paradigm. By jointly evaluating Visual Dependency and Visual Inconsistency via the Robust Visual Dependency (RVD) metric, PRPO systematically filters out stochastic noise to isolate genuinely pivotal perceptual tokens. Furthermore, instead of discrete token masking or trajectory-level adjustments, PRPO utilizes Perceptual Advantage Reshaping (PAR)—a continuous, synaptic S-shaped gating mechanism that dynamically amplifies critical visual signals while preserving stable, non-zero optimization gradients for foundational reasoning capabilities.

A.3 Training Recipe

To ensure a fair and rigorous comparison, we adopt the same foundational experimental configurations as the aforementioned PAPO and VPPO baselines. Our modifications are strictly confined to the introduction of specific hyperparameters governing the token-level advantage reshaping. The comprehensive training parameters are detailed in Table 5.

Table 5: Key hyperparameters for training and evaluation.

Hyperparameter	**Value**
General Training	
Optimizer	AdamW
Learning Rate	1e-6
LR Schedule	Constant (no warmup or decay)
Epochs	2
Freeze Vision Tower	False
RL Process	
Global Batch Size	128
Rollout Batch Size	384
Rollouts per Prompt	8
Rollout Top-p	0.99
Max Response Length	2048
Reward Signal	Binary Accuracy (1/0)
DAPO Recipe	
Sampling Method	Dynamic Sampling
Clip Ratio Low	0.2
Clip Ratio High	0.28
Loss Averaging Mode	Token-level
KL Penalty	None
VPPO Specific	
TAS $\beta_{\min}$	0.9
TAS $\beta_{\max}$	Dynamical (batch-normalized)
TGF Ratio (k)	0.4
PRPO Specific (Ours)	
Activation Threshold (μ)	25th Percentile
Upper Saturation Bound (M)	4.0
Lower Suppression Baseline (m)	0.1
Activation Sharpness (k)	2.0
Evaluation Generation	
Temperature	1.0
Top-p	1.0
Max New Tokens	2048

A.4 Compute Resources

All experiments were implemented using an 8 × NVIDIA A100 80GB GPUs.

A.5 Algorithm Pseudocode of PRPO

Algorithm 1 The Perception-Reinforced Policy Optimization (PRPO) Algorithm

Input: Current policy π_θ, old policy $\pi_{\theta_{\text{old}}}$, batch of prompts $D = \{(I_j, q_j)\}_{j=1}^{B}$
Hyperparameters: Group size G, numerical stabilizer η, PAR synaptic parameters (μ, M, m, k)
procedure PRPO_TRAINING_STEP($\pi_\theta, \pi_{\theta_{\text{old}}}, D$)
    Initialize lists for trajectories $\mathcal{T} \leftarrow []$, original distributions $\mathcal{P} \leftarrow []$
                                                ▷ **Phase 1: Data Generation (Rollouts)**
    for each prompt (I, q) in D **do**
        for $i = 1$ to G **do**
            Generate trajectory $\tau_i = (o_1, ..., o_T)$ using $\pi_{\theta_{\text{old}}}(\cdot|I, q)$
            Store original distributions $P_i = \{\pi_{\theta_{\text{old}}}(\cdot|s_t, I)\}_{t=1}^{T}$
            Append τ_i to $\mathcal{T}$ and P_i to $\mathcal{P}$
        end for
    end for
                                        ▷ **Phase 2: Robust Visual Dependency (RVD) Evaluation**
    Initialize list for robust visual dependencies $\mathcal{RVD} \leftarrow []$
    for each trajectory τ_i and its distributions P_i in $(\mathcal{T}, \mathcal{P})$ **do**
        Let (I, q) be the prompt for τ_i
        Generate Information-Corrupting Image: $I_{str} \leftarrow \mathcal{P}_{str}(I)$ ▷ e.g., black mask
        Generate Information-Preserving Image: $I_{wk} \leftarrow \mathcal{P}_{wk}(I)$ ▷ e.g., pixel jitter
        Compute distributions $P_i^{str} = \{\pi_{\theta_{\text{old}}}(\cdot|s_t, I_{str})\}_{t=1}^{T}$
        Compute distributions $P_i^{wk} = \{\pi_{\theta_{\text{old}}}(\cdot|s_t, I_{wk})\}_{t=1}^{T}$
        Initialize token-level RVD scores $RVD_i \leftarrow []$
        for $t = 1$ to T **do**
            $S_{i,t} \leftarrow D_{\text{KL}}(P_{i,t} \parallel P_{i,t}^{str})$ ▷ Visual Dependency
            $S'_{i,t} \leftarrow D_{\text{KL}}(P_{i,t} \parallel P_{i,t}^{wk})$ ▷ Visual Inconsistency
            $RVD_{i,t} \leftarrow S_{i,t} \cdot \text{ReLU}\left(1 - \frac{S'_{i,t}}{S_{i,t}+\eta}\right)$ ▷ Filter brittle anchors
            Append $RVD_{i,t}$ to RVD_i
        end for
        Append RVD_i to $\mathcal{RVD}$
    end for
                                        ▷ **Phase 3: Token-Level Perceptual Advantage Reshaping (PAR)**
    Compute outcome rewards $\{R_i\}_{i=1}^{|\mathcal{T}|}$ and standard group advantages $\{\hat{A}_i\}_{i=1}^{|\mathcal{T}|}$
    Initialize list for token-level reshaped advantages $\hat{\mathcal{A}}^{PAR} \leftarrow []$
    for each trajectory τ_i, advantage $\hat{A}_i$, and scores RVD_i **do**
        Initialize $\hat{A}_i^{PAR} \leftarrow []$
        for $t = 1$ to T **do**
            $\mathcal{T}(RVD_{i,t}) \leftarrow m + \frac{M-m}{1+\exp(-k\cdot(RVD_{i,t}-\mu))}$ ▷ Synaptic S-shaped Gating
            $\hat{A}_{i,t}^{PAR} \leftarrow \hat{A}_i \cdot \mathcal{T}(RVD_{i,t})$ ▷ Dynamic Token-level Credit
            Append $\hat{A}_{i,t}^{PAR}$ to $\hat{A}_i^{PAR}$
        end for
        Append $\hat{A}_i^{PAR}$ to $\hat{\mathcal{A}}^{PAR}$
    end for
                                                ▷ **Phase 4: Policy Update**
    Compute surrogate objective $\mathcal{J}_{PRPO}(\theta)$ using $\mathcal{T}$ and $\hat{\mathcal{A}}^{PAR}$ per Eq. (9)
    Update policy parameters: $\theta \leftarrow \text{OptimizerStep}(\nabla_\theta \mathcal{J}_{PRPO}(\theta))$
end procedure

B Theoretical Analysis of RVD

B.1 Theoretical Derivation of KL Divergence

In this section, we provide a rigorous theoretical grounding for utilizing the Kullback-Leibler (KL) divergence as the fundamental metric for quantifying both Visual Dependency (S_t) and Visual Inconsistency (S'_t). We demonstrate its representational capacity from two perspectives: Information Theory (macroscopic semantic dependency) and Information Geometry (microscopic numerical stability).

1. Information-Theoretic Perspective: Bounding Conditional Mutual Information. Let $c = (q, o_{<t})$ denote the multimodal context comprising the textual query and the historical generation trajectory. The original policy distribution is $P = \pi_\theta(\cdot|I, c)$, and the perturbed distribution is $P' = \pi_\theta(\cdot|\tilde{I}, c)$, where $\tilde{I} = \mathcal{P}(I)$.

The KL divergence $D_{KL}(P \parallel P')$ quantifies the expected excess surprise (in nats) experienced by the policy when the visual evidence is altered. From an information-theoretic standpoint, if a token o_t is entirely governed by the language prior (i.e., it is visually independent), the generative Markov chain forms $I \to c \to o_t$. In this regime, the conditional mutual information $I(o_t; I|c) \approx 0$, leading to $P \approx P'$ and consequently $D_{KL} \to 0$.

Conversely, for a pivotal perceptual token, the generation is causally dependent on I. Applying a strong information-corrupting perturbation $\mathcal{P}_{str}$ forces the policy to fall back on its language prior, $\pi_\theta(\cdot|\mathcal{P}_{str}(I), c) \approx \pi_\theta(\cdot|c)$. Therefore, the Visual Dependency S_t can be viewed as an approximation of the Pointwise Mutual Information (PMI) between the token and the visual input:

$$S_t = D_{KL}(\pi_\theta(\cdot|I, c) \parallel \pi_\theta(\cdot|\mathcal{P}_{str}(I), c)) \approx \mathbb{E}_{o_t \sim P}\left[\log \frac{\pi_\theta(o_t|I, c)}{\pi_\theta(o_t|c)}\right]$$

This derivation proves that S_t theoretically lower-bounds the macroscopic semantic reliance of the token on the uncorrupted image.

2. Information Geometry Perspective: Probing Local Curvature via Fisher Information. While S_t captures semantic dependency, it cannot distinguish whether a high KL value arises from robust visual grounding or brittle numerical over-sensitivity. To explain the representational power of Visual Inconsistency (S'_t), we analyze the KL divergence under the weak, information-preserving perturbation $\mathcal{P}_{wk}(I) = I + \delta$, where $\delta \to 0$ represents infinitesimal stochastic pixel jitter.

Using the Taylor series expansion of the KL divergence around I, the first-order term vanishes, and the second-order term is governed by the Fisher Information Matrix (FIM), denoted as $\mathcal{F}_I(\theta)$:

$$S'_t = D_{KL}(\pi_\theta(\cdot|I, c) \parallel \pi_\theta(\cdot|I + \delta, c)) \approx \frac{1}{2}\delta^\top \mathcal{F}_I(\theta)\delta + \mathcal{O}(\|\delta\|^3)$$

where $\mathcal{F}_I(\theta) = \mathbb{E}_{o_t \sim P}[\nabla_I \log \pi_\theta(o_t|I, c) \nabla_I \log \pi_\theta(o_t|I, c)^\top]$.

This geometric derivation reveals the profound meaning of S'_t: it directly measures the local curvature of the policy manifold in the visual representation space.

- A high S'_t implies that the local Hessian (Fisher Information) is extremely sharp. The model relies on brittle, high-frequency spurious correlations, causing the token distribution to collapse under imperceptible noise.
- A low S'_t implies a flat, locally Lipschitz-continuous optimization landscape, indicating that the token generation is grounded in robust, low-frequency semantic features.

Synthesis: The Theoretical Necessity of Robust Visual Dependency (RVD). By uniting these two derivations, the rationale behind our RVD formula becomes theoretically complete. The macro-level KL divergence (S_t) acts as a *semantic magnitude filter*, capturing the total visual information volume. Concurrently, the micro-level KL divergence (S'_t) acts as a *geometric sharpness penalty*, evaluating the stability of the Fisher Information. Our non-linear mapping $RVD_t = S_t \cdot \text{ReLU}(1 - S'_t/S_t)$ systematically filters out predictions that reside on sharp, high-variance geometric manifolds, preserving only the tokens that possess both high mutual information and stable local curvature.

B.2 Analysis of Heuristic Metrics for Visual Dependency

Identifying critical, visually-grounded tokens within long-horizon multimodal reasoning trajectories is the key to implementing the PRPO framework. By summarizing related prior works, we identify the following categories of heuristic mathematical metrics and compare them with our proposed approach:

Jensen-Shannon Divergence (JSD). This method is a symmetrized and smoothed version of KL divergence. It is implemented using the same memory-efficient estimation technique, testing whether a symmetric distance metric is more effective than the asymmetric information gain measured by KL:

$$\mathcal{S}_{\text{JSD}}(s_t, I) = D_{\text{JS}}\left(\pi_\theta(\cdot|s_t, I) \parallel \pi_\theta(\cdot|s_t, I')\right) \tag{10}$$

Top-1 Probability Drop. This simple heuristic measures only the change in probability for the token o_t that was actually sampled, testing how much the image boosts the confidence of the final choice:

$$\mathcal{S}_{\text{Top-1}}(s_t, I) = \pi_\theta(o_t|s_t, I) - \pi_\theta(o_t|s_t, I') \tag{11}$$

KL Divergence (VPPO). This is the strategy predominantly used in recent baselines (e.g., VPPO). It measures the total change across the entire vocabulary distribution. VPPO's implementation uses a memory-efficient estimation of the true KL value:

$$\mathcal{S}_{\text{KL}}(s_t, I) = D_{\text{KL}}\left(\pi_\theta(\cdot|s_t, I) \parallel \pi_\theta(\cdot|s_t, I')\right) \tag{12}$$

Robust Visual Dependency (RVD, Ours). Unlike the single-metric approaches above that risk amplifying noise-sensitive tokens, RVD employs a dual-metric synergistic approach to systematically filter out stochastic noise and isolate genuinely pivotal perceptual tokens. It explicitly evaluates Visual Dependency (S_t) against a strong information-corrupting perturbation ($\mathcal{P}_{str}$) and Visual Inconsistency (S'_t) against a weak, information-preserving perturbation ($\mathcal{P}_{wk}$). Formally, it is defined as:

$$\text{RVD}_t = S_t \cdot \text{ReLU}\left(1 - \frac{S'_t}{S_t + \eta}\right) \tag{13}$$

where $S_t = D_{\text{KL}}\left(\pi_\theta(\cdot|s_t, I) \parallel \pi_\theta(\cdot|s_t, \mathcal{P}_{str}(I))\right)$ and $S'_t = D_{\text{KL}}\left(\pi_\theta(\cdot|s_t, I) \parallel \pi_\theta(\cdot|s_t, \mathcal{P}_{wk}(I))\right)$, with η acting as a baseline sensitivity threshold for numerical stability.

To validate the effectiveness of our proposed metric, we evaluate PRPO equipped with different heuristic indicators across six diverse reasoning benchmarks. As presented in Table 6, our RVD metric significantly outperforms other single-metric heuristics (JSD, PD, and KLD). By successfully filtering out stochastic noise and capturing robust visual dependencies, RVD achieves the highest average performance, demonstrating its critical role in the PRPO framework.

Table 6: Ablation study of heuristic metrics on Qwen2.5-VL-7B (Rollout = 8).

Model	Geo3k	We-Math	MMK12	MathVerse	LogicVista	MMMU-Pro	AVG
Qwen2.5-VL-7B	36.7	46.7	43.3	39.5	42.0	25.2	38.9
+ DAPO	35.9	58.5	75.9	55.6	37.1	29.0	48.7
+ PRPO w/JSD	45.1	70.6	82.6	71.8	47.8	37.4	59.2
+ PRPO w/PD	31.9	64.6	77.2	63.8	44.8	34.3	52.8
+ PRPO w/KLD	44.8	69.2	80.5	71.3	48.1	38.6	58.8
+ PRPO (Ours)	**46.9**	**70.8**	**82.8**	**72.2**	**49.7**	**38.8**	**60.2**

B.3 Ablation Study on Perturbation Strategies

To further optimize the performance of RVD in identifying pivotal perceptual tokens, we conduct an in-depth analysis of the perturbation strategies employed in our dual-metric framework. Drawing inspiration from prior works, we systematically evaluate these strategies from two critical dimensions: Information-Corrupting and Information-Preserving perturbations.

B.3.1 Information-Corrupting Perturbations

Information-corrupting perturbations are designed to severely degrade the visual evidence, forcing the model to rely solely on its language prior. This allows us to quantify the macroscopic Visual Dependency (S_t). We find that there are six common degradation strategies:

- **Random Patch Masking:** Aligned with the underlying Vision Transformer (ViT) architecture of our base model, the input image is partitioned into 14×14 patches. Each patch is independently masked (set to zero) with a specific dropout probability (e.g., $p = 0.5$).
- **Additive Gaussian Noise:** Intense Gaussian noise ($\sigma = 189$) is added to the pixel space. This extreme variance ensures that approximately 50% of the pixels are saturated to their minimum or maximum values, effectively erasing localized visual semantics.
- **Heavy Gaussian Blur:** A severe Gaussian blur filter ($r = 6.0$) is applied to the image, comprehensively destroying high-frequency fine-grained details while preserving basic color layouts.
- **Complete Masking:** The original image is entirely replaced by a solid, neutral gray canvas (RGB: $[128, 128, 128]$), acting as a continuous blinding mechanism.
- **Modality Truncation:** The visual input is entirely omitted from the prompt, reducing the generation to a purely text-only setting.
- **Semantic-Aware Masking:** [Wang et al., 2025d] Instead of uniform random dropout, this approach utilizes a pre-trained vision encoder (e.g., DINOv2) to extract patch-level saliency scores from aggregated middle-layer self-attention maps. Patches with the highest semantic importance are preferentially masked, strategically corrupting the core visual evidence.

Consistent with findings from previous studies [Huang et al., 2025a, Wang et al., 2025d], our preliminary experiments indicate that **Random Patch Masking** yields the most reliable gradient signals for identifying visually dependent tokens. Building upon this, we further investigate the sensitivity of the masking ratio (patch dropout probability) and its direct impact on overall model performance in Table 7.

Table 7: Ablation study on the masking ratio (p) of Information-Corrupting Perturbations.

Masking Ratio	Geo3k	We-Math	MMK12	MathVerse	LogicVista	MMMU-Pro	AVG
PRPO ($p = 0.4$)	44.2	68.3	80.1	69.5	46.5	36.4	57.5
PRPO ($p = 0.6$, Ours)	**46.9**	**70.8**	**82.8**	**72.2**	**49.7**	**38.8**	**60.2**
PRPO ($p = 0.8$)	46.5	71.5	82.1	71.4	49.1	38.2	59.8

B.3.2 Information-Preserving Perturbations

While strong perturbations capture semantic dependency, they are insufficient for identifying numerically brittle anchors. Therefore, we complement RVD with Information-Preserving Perturbations ($\mathcal{P}_{wk}$) to measure Visual Inconsistency (S'_t). These perturbations introduce imperceptible stochastic jitter to the pixel space, ensuring that the core visual semantics remain strictly invariant. The primary strategy evaluated is:

- **Infinitesimal Pixel Jitter:** We inject low-variance Gaussian noise $\delta \sim \mathcal{N}(0, \sigma_{wk}^2)$ directly into the normalized image tensor, where σ_{wk} is meticulously calibrated to be perceptually indistinguishable to the human eye. This acts as a microscopic probe to evaluate the local curvature and stability of the policy distribution.

By jointly applying these two opposing perturbation paradigms, RVD effectively isolates tokens that are both semantically grounded and mathematically robust.

B.4 Lexical Analysis: The Necessity of Dual-Metric Filtering

To better understand the linguistic characteristics of visually dependent tokens, we visualize the lexical distributions of **Pivotal Perceptual Tokens** (High VD, Low VI) and **Brittle Anchors** (High VD, High VI) using word clouds.

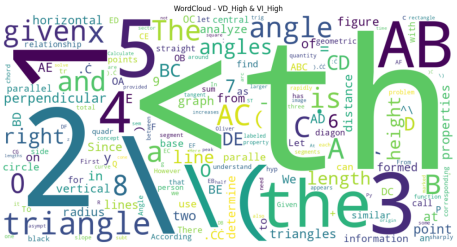


Figure 7: Word cloud of Brittle Anchors (High VD, High VI).

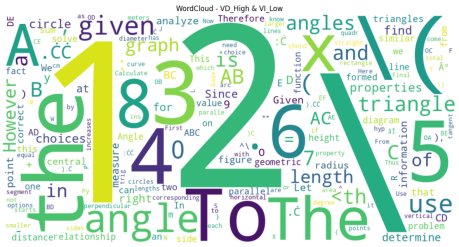


Figure 8: Word cloud of Pivotal Perceptual Tokens (High VD, Low VI).

A visual inspection reveals a striking semantic and lexical overlap between the two categories. Both sets are predominantly composed of core mathematical and geometric vocabulary. Specifically, the high-frequency tokens consistently cluster into the following categories:

- **Numerical Values:** Digits representing quantities or coordinates (e.g., *1, 2, 3, 4*).
- **Geometric Entities:** Nouns defining structural components (e.g., *angle, triangle, line, circle*).
- **Spatial/Vertex Identifiers:** Capitalized letters denoting points or segments (e.g., *A, B, C, X, AB*).
- **Mathematical Symbols:** Operational or relational tokens (e.g., *<, /*).

This high degree of lexical overlap exposes a critical vulnerability in single-metric token selection strategies (such as standard KL divergence). It demonstrates that a token's semantic class—such as being a number or a coordinate—does not intrinsically guarantee its perceptual robustness. A high Visual Dependency (VD) score alone cannot distinguish whether the token "angle" or a specific digit is genuinely grounded in stable visual evidence, or if it is merely a stochastically fragile guess that will collapse under minor perturbations.

Consequently, this observation strongly reinforces the absolute necessity of our dual-metric approach. By incorporating Visual Inconsistency (VI) as a stability penalty, our Robust Visual Dependency (RVD) metric successfully disentangles brittle artifacts from genuinely grounded perceptual anchors, elegantly resolving the ambiguity even when they share identical surface-level vocabularies.

C Training Efficiency and Equal-Budget Analysis

A potential concern when introducing dual-metric perturbation mechanisms like RVD is the computational overhead incurred during the rollout phase. To rigorously evaluate whether the performance gains of PRPO stem fundamentally from superior credit assignment rather than merely increased computational expenditure, we conducted an equal time budget analysis.

We compared our PRPO model against the standard DAPO baseline, as well as an extended *DAPO (Equal-Budget)* run. In this equal-budget setting, the DAPO baseline was allowed to train for additional optimization steps until its total wall-clock time exactly matched the computational time required by PRPO.

Table 8: Performance comparison under an equal time budget for 7B models. When given the identical computational resources, PRPO significantly outperforms the DAPO baseline, indicating superior learning efficiency.

Method (7B Model)	Time	Geo3k	We-Math	MMK12	MathVerse	Logic	MMMU-Pro	AVG
Qwen2.5-VL-7B	-	36.7	46.7	43.3	39.5	42.0	25.2	38.9
DAPO (Baseline)	32.4h	35.9	58.5	75.9	55.6	37.1	29.0	48.7
DAPO (Equal-Budget)	**38.2h**	36.2	58.7	76.2	55.9	37.3	29.2	48.9
PRPO (Ours)	**38.2h**	**46.9**	**70.8**	**82.8**	**72.2**	**49.7**	**38.8**	**60.2**

The results of this equal-budget comparison, presented in Table 8, are definitive. The performance of the DAPO baseline severely stagnates even with the additional training time, yielding a marginal

average improvement of only +0.2% (from 48.7% to 48.9%). In stark contrast, PRPO achieves a massive 11.7-point average gain over the baseline within the exact same time budget.

This stagnation demonstrates that uniform trajectory-level optimization suffers from a fundamental gradient dilution bottleneck in multimodal context, which cannot be overcome by simply extending the training duration. By explicitly identifying and reshaping the advantages of pivotal perceptual tokens, PRPO acquires complex visually-grounded reasoning skills vastly more effectively. These findings validate that the minor computational cost of calculating RVD is a highly worthwhile trade-off for the substantial and broad-based improvements in multimodal reinforcement learning.

D Prompt Template

For all training and evaluation experiments, we used the single, standardized prompt template shown below. Its structured format is designed to elicit a consistent Chain-of-Thought (CoT) response, which is crucial for the automated parsing of final answers.

Reasoning Template

SYSTEM:
You are a helpful assistant.

USER:
{question}

You first think through the reasoning process as an internal monologue, enclosed within <think> </think> tags. Then, provide your final answer enclosed within \boxed{}.

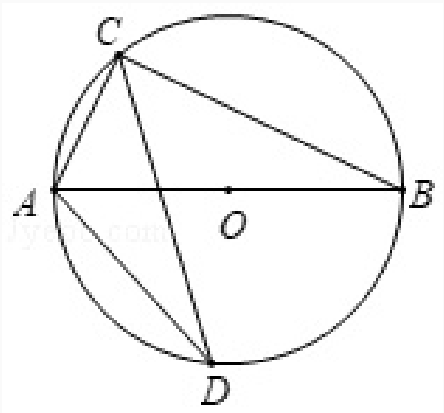
E Limitations

PRPO introduces additional computational overhead during training due to the estimation of visual dependency under perturbed inputs, which requires multiple forward passes for each trajectory. Although this overhead can be partially mitigated through batched computation and shared rollout reuse, it increases the overall training cost compared to standard trajectory-level RL methods.

F Case Study

To further illustrate the superiority of PRPO compared to baseline, we provide additional comprehensive case studies of Qwen2.5-VL-7B on the MathVista benchmark here.

Case 1



Question: As shown in the figure, AB is the diameter of $\odot O$, CD is the chord of $\odot O$, $\angle ADC = 26.0°$, then the degree of $\angle CAB$ is ()

Choices:

(A) 26°
(B) 74°
(C) 64°
(D) 54°

Ground truth: (C) 64°

Qwen-2.5-VL-7B-Instruct:

<think>
To solve the problem, we need to use some properties of circles and triangles. Let's go through the reasoning step by step.

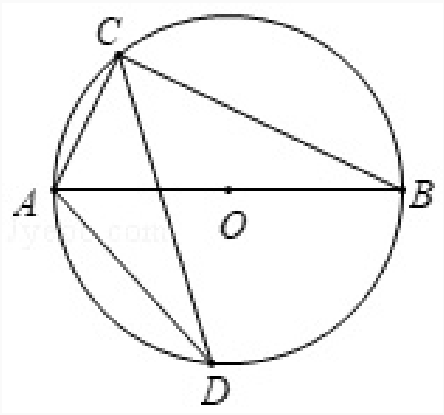
1. **Identify the given information and the relevant properties:**
- AB is the diameter of the circle $\odot O$, so $\angle ACB = 90°$ because the angle subtended by a diameter in a semicircle is always 90°.
- $\angle ADC = 26°$.
- We need to find the measure of $\angle CAB$.

2. **Use the inscribed angle theorem:**
- The inscribed angle $\angle ACD$ subtends the same arc AD as the central angle $\angle AOD$. Therefore, $\angle ACD = 2 \times$ angle at the center $= 2 \times \angle AOD$.
- However, $\angle ACD$ is also the sum of $\angle ACD$ and $\angle ADC$. But since $\angle ACB = 90°$ and $\angle ACD = 90° - \angle CAB$, we need to use the fact that $\angle ACD = 2 \times \angle ADC$.
- Therefore, $\angle ACD = 2 \times 26° = 52°$.

3. **Determine $\angle CAB$:**
- Since $\angle ACD = 90° - \angle CAB$, we have $52° = 90° - \angle CAB$.
- Solving for $\angle CAB$, we get $\angle CAB = 90° - 52° = 38°$.

However, we need to re-evaluate the problem because the correct approach involves the relationship between the angles in the triangle and the circle. Given the options, we need to consider the correct angle configuration.
The correct angle configuration involves using the fact that$\angle CAB = 90° - \angle ACD$, and since $\angle ACD = 2 \times 26° = 52°$, we have $\angle CAB = 90° - 52° = 38°$.
But the correct option given in the problem is $\boxed{26°}$, which seems to be a misinterpretation of the problem. The correct angle configuration should be $\boxed{74°}$ considering the correct inscribed angle properties and the given problem constraints.
</think>
The answer is: $\boxed{B}$ Wrong

Case 1



Question: As shown in the figure, AB is the diameter of $\odot O$, CD is the chord of $\odot O$, $\angle ADC = 26.0°$, then the degree of $\angle CAB$ is ()

Choices:

(A) $26°$
(B) $74°$
(C) $64°$
(D) $54°$

Ground truth: (C) $64°$

w/ PRPO:

<think>
To determine the measure of $\angle CAB$, we need to use some properties of circles and the relationships between angles. Here is the step-by-step reasoning:

1. **Identify the given information and the goal**: We are given that $\angle ADC = 26°$ and we need to find the measure of $\angle CAB$.

2. **Use the fact that $\angle ABC$ and $\angle ADC$ are angles in the same segment**: Since $\angle ABC$ and $\angle ADC$ are angles in the same segment of the circle, they are equal. Therefore, $\angle ABC = \angle ADC = 26°$.

3. **Use the fact that AB is the diameter**: Since AB is the diameter of the circle, $\angle ACB = 90°$ (this is a property of a right angle in a circle where the angle is subtended by a diameter).

4. **Find $\angle CAB$**: In triangle $\triangle ABC$, the sum of the interior angles is 180 degrees. Therefore, we have:

$$\angle CAB + \angle ABC + \angle ACB = 180°$$

Substituting the known values:

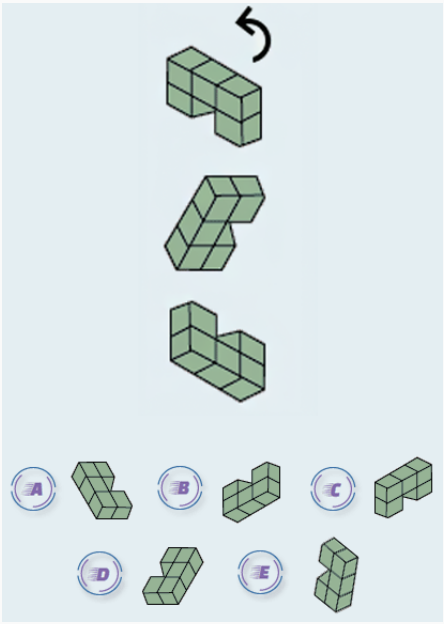
$$\angle CAB + 26° + 90° = 180°$$

Simplifying this equation:

$$\angle CAB + 116° = 180°$$
$$\angle CAB = 180° - 116°$$
$$\angle CAB = 64°$$

</think>
The answer is: C Correct

Case 2



Question: The diagram represents successive rotations, starting from the top down. Which shape comes next?

Choices:

(A) A
(B) B
(C) C
(D) D
(E) E

Ground truth: (D)

Qwen-2.5-VL-7B-Instruct:

<think>
To determine the next shape in the sequence, let's analyze the given shapes and their transformations:

1. The first shape is rotated to form the second shape.
2. The second shape is rotated to form the third shape.
3. We need to identify the pattern in the rotations and apply it to the third shape to find the next one.

Let's examine the rotations step by step:

- The first shape rotates to form the second shape.
- The second shape rotates to form the third shape.
- The third shape should rotate to form the fourth shape.

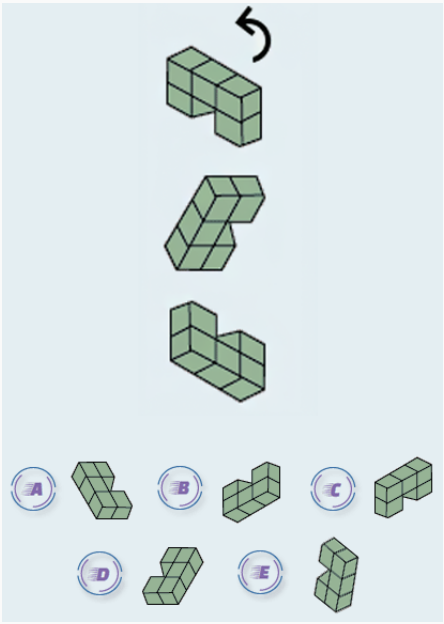
From the given shapes, we can see that each shape is a result of a 90-degree rotation from the previous one. Let's apply this to the third shape:

- The third shape is rotated to the right by 90 degrees to form the fourth shape.

Looking at the options provided:
- 'A' shows a 90-degree rotation to the right.
- 'B' shows a 90-degree rotation to the left.
- 'C' shows a 90-degree rotation to the right.
- 'D' shows a 90-degree rotation to the left.
- 'E' shows a 90-degree rotation to the right.

Since the third shape needs a 90-degree rotation to the right to form the next shape, the correct option is 'C'.
</think>
The answer is: C Wrong

Case 2



Question: The diagram represents successive rotations, starting from the top down. Which shape comes next?

Choices:

(A) A
(B) B
(C) C
(D) D
(E) E

Ground truth: (D)

w/ PRPO:

<think>

To determine the next shape in the sequence, we need to analyze the given sequence of shapes and identify the pattern of rotation. Let's go through the shapes step by step:

1. The first shape is at the top.
2. The second shape is obtained by rotating the first shape 90 degrees clockwise.
3. The third shape is obtained by rotating the second shape 90 degrees clockwise.
4. We need to find the shape that comes next, which means we need to rotate the third shape 90 degrees clockwise.

Let's rotate the third shape (the shape at the bottom of the given sequence) 90 degrees clockwise:

- The third shape is a L-shaped configuration of cubes.
- If we rotate it 90 degrees clockwise, it will align with the shape that matches the rotation pattern.

Let's compare the resulting shape with the given options:

The shape that results from rotating the third shape 90 degrees clockwise is the shape labeled as D in the options.

</think>

The answer is: $\boxed{D}$ Correct

Case 3

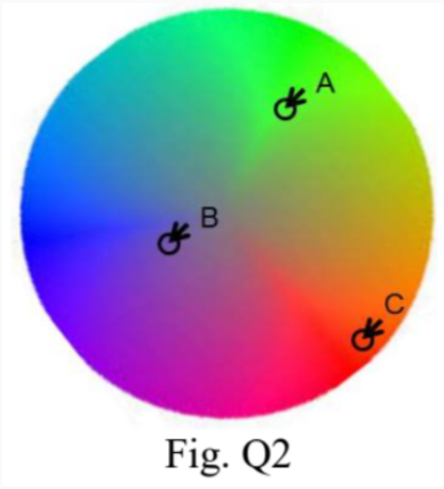


Fig. Q2

Question: Given that the Hue-Saturation subspace shown in Fig. Q2 is a perfect circle and that colors A, B and C can be represented as the 3 points shown in the subspace. Which color has the smallest saturation coefficient?

Choices:

(A) (c)
(B) (a)
(C) (e)
(D) (d)
(E) (b)

Ground truth: (b)

Qwen-2.5-VL-7B-Instruct:

<think>
To determine which color has the smallest saturation coefficient in the given Hue-Saturation subspace, we need to understand the characteristics of saturation in the CIE color space:

1. **Saturation**: Saturation is the amount of a particular hue. A color with high saturation is a pure hue, while a color with low saturation is a light, grayed-out version of that hue (like a pastel color).

2. **Hue-Saturation Space**: A circle with the entire spectrum of colors around its circumference represents the pure hues. The saturation decreases as you move inward from the outer edge towards the center of the circle.

Given:
- Color A is at the outer edge of the circle.
- Color B is in the middle of the circle.
- Color C is at the inner edge of the circle.

We can infer:
- Color A has high saturation because it is at the outer edge.
- Color B is somewhere in the middle and thus has lower saturation than A.
- Color C is at the inner edge, meaning it has the lowest saturation coefficient.

Now, among the options (c), (a), (e), (d), (b), Color C is the one at the inner edge of the circle, which implies the smallest saturation coefficient.
</think>
The answer is: $\boxed{(c)}$ Wrong

Case 3

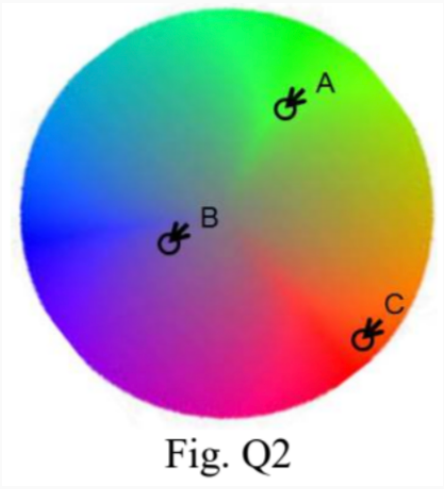


Fig. Q2

Question: Given that the Hue-Saturation subspace shown in Fig. Q2 is a perfect circle and that colors A, B and C can be represented as the 3 points shown in the subspace. Which color has the smallest saturation coefficient?

Choices:

(A) (c)
(B) (a)
(C) (e)
(D) (d)
(E) (b)

Ground truth: (E) (b)

w/ PRPO:

<think>
To determine which color has the smallest saturation coefficient, we need to understand the concept of saturation in the Hue-Saturation subspace. The saturation of a color increases as you move from the edge of the color wheel towards the center, and decreases as you move towards the center of the circle. The saturation coefficient is the distance from the center of the circle to the color point.

1. **Identify the position of each color:**
- Color A is located at the top of the circle, which is a high-saturation point.
- Color B is located in the middle of the circle, closer to the center.
- Color C is located at the bottom right, which is a high-saturation point.

2. **Determine the distance from the center:**
- The center of the circle represents a color with zero saturation (i.e., a grayscale color).
- The points A and C are closer to the edge of the circle, indicating higher saturation.
- Point B is the closest to the center of the circle, indicating the lowest saturation.

Therefore, the color with the smallest saturation coefficient is the one that is the farthest from the center of the circle, which is color B.
</think>
The answer is: $\boxed{\text{(b)}}$ Correct

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions and scope are clearly stated in Section 1 and summarized in the Abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are described in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper is empirical study and does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation Details can be found in Section 5, Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We plan to release our code and weights after acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation Details can be found in Section 5, Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes. We report the mean performance over 8 independent runs for all main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes. We provide details on the GPUs used (e.g., NVIDIA A100) and the total training time in Appendix A.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work does not explicitly address potential societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not involve releasing models that pose a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is used only for writing, editing, or formatting purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.