Not All Tokens See Equally: Perception-Grounded Policy Optimization for Large Vision-Language Models

Anonymous ACL submission

Abstract

While Reinforcement Learning from Verifiable Rewards (RLVR) has advanced reasoning in Large Vision-Language Models (LVLMs), prevailing frameworks suffer from a foundational methodological flaw: by distributing identical advantages across all generated tokens, these methods inherently dilute the learning signals essential for optimizing the critical, visually-grounded steps of multimodal reasoning. To bridge this gap, we formulate *Token Visual Dependency*, quantifying the causal information gain of visual inputs via the Kullback-Leibler (KL) divergence between visual-conditioned and text-only predictive distributions. Revealing that this dependency is highly sparse and semantically pivotal, we introduce Perception-Grounded Policy Optimization (PGPO), which is a novel fine-grained credit assignment framework that dynamically reshapes advantages at the token level. Through a threshold-gated, mass-conserving mechanism, PGPO actively amplifies learning signals for visually-dependent tokens while suppressing gradient noise from linguistic priors. Extensive experiments based on the Qwen2.5-VL series across seven challenging multimodal reasoning benchmarks demonstrate that PGPO boosts models by 18.7% on average. Both theoretical and empirical analyses confirm that PGPO effectively reduces gradient variance, prevents training collapse, and acts as a potent regularizer for robust, perception-grounded multimodal reasoning.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) (Zhang et al., 2025)—especially online methods such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024)—has substantially improved reasoning performance in text-oriented Large Language Models (LLMs). Motivated by these gains, recent studies have extended RLVR to Large Vision-Language Models (LVLMs), with progress mainly driven by improvements in training data design (Li et al., 2025a; Liang et al., 2025; Liu et al., 2025; Yao et al., 2025; Chen et al., 2025a), reward construction (Shen et al., 2025; Xia et al., 2025; Wang et al., 2025c), and optimization strategy refinement (Wang et al., 2025a; Zhao et al., 2025; Huang et al., 2026).

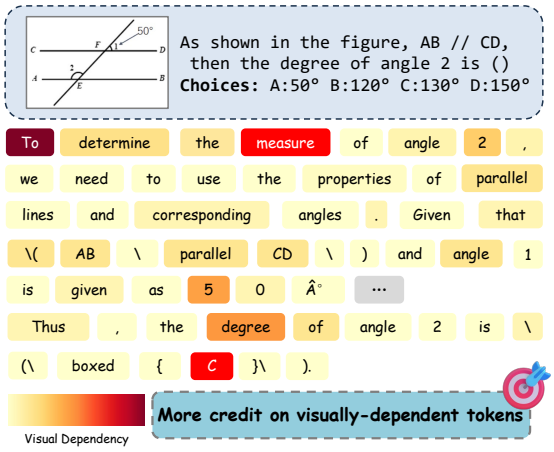


Figure 1: Unlike standard uniform credit assignment, our proposed method dynamically allocates higher reinforcement learning advantage to pivotal tokens that heavily rely on visual perception.

Although existing RLVR studies for LVLMs report notable progress, most optimization pipelines still underemphasize fine-grained visual perception. This creates a key methodological weakness: multimodal reasoning depends on precise perception (Xiao et al., 2025; Wang et al., 2025d), which serves as the basis for reliable downstream inference. In practice, this demand for visual grounding is not uniformly distributed across the generated text. As illustrated in Figure 1, only a critical minority of tokens (e.g., specific geometric entities or numerical values extracted from an image) exhibit high visual dependency. Despite this inherent token-level difference, standard GRPO relies on a coarse-grained credit assignment mechanism, uni-

formly distributing a single sequence-level, group-normalized advantage across all tokens in the generated trajectory. In the context of multimodal reasoning, this poses a foundational flaw: by treating all tokens equally, it fundamentally dilutes the learning signals essential for optimizing those pivotal, perception-grounded reasoning steps.

To address this limitation, it is imperative to quantify the causal contribution of the visual modality to each individual reasoning step. We formalize this concept as *Token Visual Dependency*, measured by the Kullback-Leibler (KL) divergence between the model's predictive distributions with and without visual conditioning. This metric captures the Bayesian surprise (Itti and Baldi, 2009)—the specific information gained from the visual context. In Section 3, we find this dependency exhibits a highly sparse distribution, demonstrating that visually-grounded reasoning is driven by a critical minority of tokens. Furthermore, it reliably isolates semantic visual grounding by assigning higher values to image-dependent visual anchors and truthful concepts.

To this end, we propose **Perception-Grounded Policy Optimization (PGPO)**, a novel token-level credit assignment framework that integrates threshold-gated, outcome-preserving advantage reallocation into multimodal RLVR. PGPO redistributes advantages based on normalized visual dependency scores: it suppresses the gradient noise from low-dependency tokens and actively boosts the learning signal for perceptually pivotal tokens. To ensure stable optimization, PGPO applies a sum-preserving normalization that keeps the overall advantage mass consistent with the original, guaranteeing a zero mean for intra-group advantages.

To validate the effectiveness of PGPO, we conduct extensive experiments across seven challenging multimodal reasoning benchmarks. Based on Qwen2.5-VL (Bai et al., 2025b) series models, PGPO achieves state-of-the-art performance comparison with previous methods. Crucially, by performing fine-grained, token-level advantage reshaping, PGPO reduces gradient variance, accelerates convergence, and increases visual dependency, thereby acting as a potent regularizer for robust, perception-grounded multimodal reasoning.

Our contributions can be summarized as follows:

- We formulate token visual dependency to quantify the causal impact of visual information, and empirically validate its sparse distribution and correlation with visual semantics.
- We propose PGPO, a fine-grained credit assignment framework that utilizes token-level advantage reshaping to amplify learning signals for visually-dependent tokens while suppressing noise.
- Extensive empirical evaluations and theoretical proofs demonstrate that PGPO achieves state-of-the-art performance on diverse multimodal benchmarks, covering mathematical, geometric, logical, multi-discipline reasoning and general VQA.

2 Related Work

Multimodal Reasoning. To narrow the remaining performance gap of LVLMs on complex reasoning tasks (Wang et al., 2024b; Dong et al., 2025), recent work has extended core reinforcement learning methods such as PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), and DAPO (Yu et al., 2025) from text settings to multimodal training (Guo et al., 2025; Bai et al., 2025b; Hurst et al., 2024). Most existing multimodal RL pipelines, however, improve components around the optimizer rather than the token-level optimization mechanism itself. Prior work mainly falls into data-oriented designs (Bai et al., 2025c; Li et al., 2025a; Chen et al., 2025c; Wei et al., 2025b; Chen et al., 2025b) and reward-oriented designs that inject perception-related supervision (Wang et al., 2025d; Ma et al., 2025), together with rollout-level changes and external visual tool usage (Liu et al., 2025; Wang et al., 2025a). Huang et al. (2026) improves the optimization of visually dependent tokens through binary gradient filtering, but it lacks dynamic token-level advantage allocation. Even with these advances, many methods still assign identical training signals to every generated token, which is mismatched with the fine-grained nature of multimodal reasoning.

Credit Assignment. Credit assignment (Sutton and Barto, 1998; Arumugam et al., 2021; Zhou et al., 2020) is a central problem in reinforcement learning: determining which earlier actions should receive responsibility for later outcomes. This issue becomes more severe in RLVR for LLMs, where sparse terminal rewards must be traced back to token-level decisions across long reasoning trajectories. To mitigate this challenge, recent work

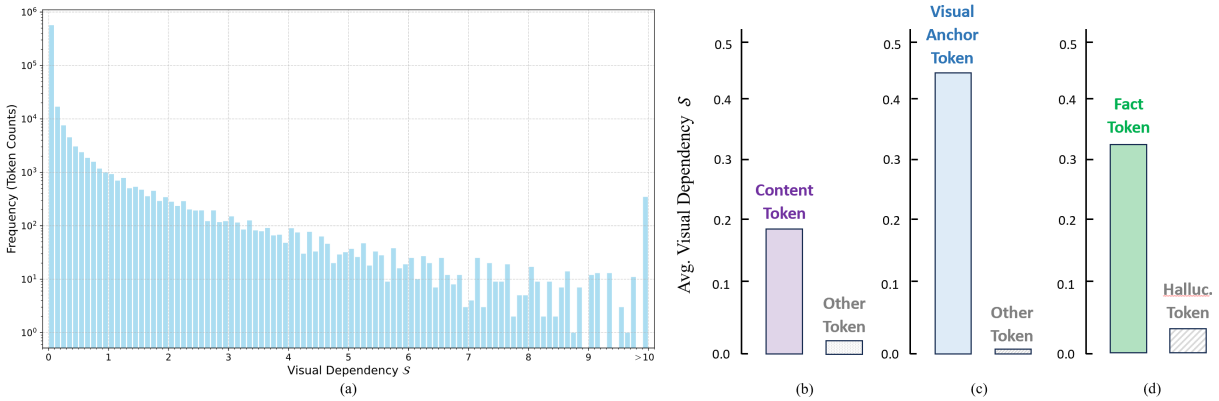


Figure 2: Empirical analysis results of visual dependency. (a) The skewed distribution of token-level visual dependency. (b,c,d) Average visual dependency comparison between specific category tokens and others.

(Wang et al., 2025b; Wei et al., 2025a; Li et al., 2025b) explores fine-grained attribution for more precise advantage estimation. Yet, principled token-level credit redistribution for perception-grounded reasoning in LVLMs remains underexplored.

3 Preliminaries

To study the perception in multimodal reasoning, this section first formalizes *Token Visual Dependency* via Kullback-Leibler (KL) divergence to quantify the causal information gain from visual inputs at each reasoning step. We then empirically demonstrate that this visual dependency exhibits a highly sparse distribution, strongly correlating with semantic visual anchors and factual truthfulness. Together, these theoretical and empirical findings expose the fundamental limitations of uniform sequence-level rewards, directly motivating the core design of our Perception-Grounded Policy Optimization (PGPO) approach.

3.1 Token Visual dependency

We characterize token-level visual dependency as the incremental information contributed by the image context, inspired by Huang et al. (2026). We measure it using the Kullback-Leibler (KL) divergence between the policy's next-token distribution with visual conditioning and the corresponding distribution when visual input is removed, thereby capturing the image-induced distributional change.

Let I denote the image, q the textual query, and o the generated response. For state $s_t = (q, o_{<t})$, we define the visual dependency score at step t as:

$$\mathcal{S}(s_t, I) := D_{\mathrm{KL}}\left(\pi_\theta(\cdot|s_t, I) \parallel \pi_\theta(\cdot|s_t)\right). \quad (1)$$

A larger $\mathcal{S}$ implies that visual evidence is indispensable for predicting the token at step t, indicating a strongly perception-grounded reasoning step. By quantifying Bayesian surprise—the information gained when updating a language prior to a visually-grounded posterior—KL divergence uniquely isolates the causal intervention of the visual modality on each generated token. Furthermore, its strict non-negativity, mathematically guaranteed by Gibbs' inequality. These favorable theoretical properties yield a robust credit assignment signal for reinforcement learning. More theoretical analyses can be found in Appendix E, which is fundamentally grounded in the principles of the information theory.

3.2 Empirical Analysis of Visual Dependency

Having established the theoretical formulation of token-level visual dependency $\mathcal{S}$, we proceed to empirically validate its behavior in multimodal reasoning tasks. In this section, we present a comprehensive analysis of its distributional sparsity and categorize tokens based on their roles in the reasoning process to compare their average $\mathcal{S}$ on Qwen2.5-VL-3B. Together, these empirical findings substantiate $\mathcal{S}$ as a interpretable metric for guiding policy optimization in visual reasoning. Detailed settings are provided in Appendix D.

Distributional Sparsity of Visual Dependency. We further examine the empirical pattern of token-level visual dependency through the proposed metric. Specifically, we run inference on the vision-dominant portion of MathVerse (Zhang et al., 2024) and compute $\mathcal{S}$ for each generated token in every trajectory. Figure 2(a) shows that, under a logarithmic view, token frequency decreases rapidly as $\mathcal{S}$ becomes larger, revealing a strongly long-tailed distribution. This observation indicates that

visually grounded reasoning is driven by only a limited subset of tokens. In addition, Figure 2(b) highlights the semantic relevance of these tokens: content-bearing tokens, such as numbers, geometric concepts, verbs, and adjectives, attain noticeably higher mean $\mathcal{S}$ values than functionally less informative tokens. Therefore, assigning the same optimization signal to every token is suboptimal, because it spreads credit to many steps that contribute little to actual visual reasoning.

Higher Dependency for Visual Anchors. Next, we investigate whether $\mathcal{S}$ is associated with critical visual information. We utilize generated trajectories from the MathVerse vision-dominant subset and annotate *visual anchor words* by GPT-5—defined as tokens that are impossible to properly infer without observing the image. Figure 2(c) reveals that these visual anchor tokens exhibit significantly higher average $\mathcal{S}$ compared to non-anchor tokens. This confirms that $\mathcal{S}$ effectively captures semantic visual grounding, making it a meaningful metric for credit assignment to enhance visual reasoning in RLVR.

Lower Dependency in Hallucinations. Finally, we explore the correlation between $\mathcal{S}$ and model hallucinations using the CHAIR (Rohrbach et al., 2019) benchmark. By generating image captions, we extract and categorize tokens into truthful and hallucinated. As shown in Figure 2(d), truthful tokens demonstrate higher average $\mathcal{S}$ than hallucinated ones. This suggests that hallucinated concepts are primarily driven by the model's language prior rather than actual visual input, further validating $\mathcal{S}$ as a possible proxy for factual visual reliance.

4 Methodology

Based on above empirical analysis of token visual dependency, to capitalize on fine-grained perceptual signals, we propose Perception-Grounded Policy Optimization (PGPO). It integrates token-level visual dependency with a threshold-gated, mass-conserving advantage reallocation mechanism.

4.1 Group Relative Policy Optimization

In the context of multimodal Reinforcement Learning with Verifiable Rewards (RLVR), a Large Vision-Language Model (LVLM) defines an autoregressive policy π_θ. Given a multimodal prompt (I, q) consisting of a visual input I and a textual query q, the old policy $\pi_{\theta_{old}}$ generates a group of G responses, $\{o_i\}_{i=1}^G$. A reward R_i is assigned to each complete response based on whether its final extracted answer matches the ground truth.

GRPO normalizes the outcome reward within the sampled group to compute the advantage $\hat{A}_i$ for a response o_i:

$$A_i = \frac{R_i - mean(\{R_k\}_{k=1}^G)}{std(\{R_k\}_{k=1}^G)} \tag{2}$$

This single, coarse sequence-level advantage A_i is then broadcast indiscriminately to every timestep t in the sequence. The policy π_θ is updated by maximizing the clipped surrogate objective:

$$\mathcal{L}_{GRPO}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(\rho_{i,t}(\theta) A_i, \text{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_i\Big)\Bigg] \tag{3}$$

where $\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|I,q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|I,q,o_{i,<t})}$ represents the probability ratio. However, this uniform assignment neglects the varying contribution of the visual modality of individual steps, resulting in suboptimal visual reasoning and increased gradient variance for tokens that lack visual dependency.

4.2 Token Visual Dependency Score

The magnitude of $\mathcal{S}_t$ directly quantifies the t-th token's reliance on visual grounding. However, KL divergence is inherently unbounded and typically exhibits a heavy-tailed distribution, making it unsuitable for direct credit assignment. To construct a well-behaved, bounded visual dependency score $I_t \in [0, 1]$ for downstream advantage reshaping, we apply a two-stage monotonic transformation.

First, we mitigate the skewed distribution of the raw divergence via a shifted logarithmic compression, defining the dampened signal $\tilde{\mathcal{S}}_t$ as:

$$\tilde{\mathcal{S}}_t = \log(1 + \mathcal{S}_t) \tag{4}$$

This shift guarantees non-negativity by strictly anchoring zero information gain to zero. More importantly, it prevents extreme outliers from disproportionately dominating the subsequent normalization step, thereby preserving the variance of moderate yet informative attribution signals.

Second, to project these unbounded dampened signals onto a standardized relative scale, we apply

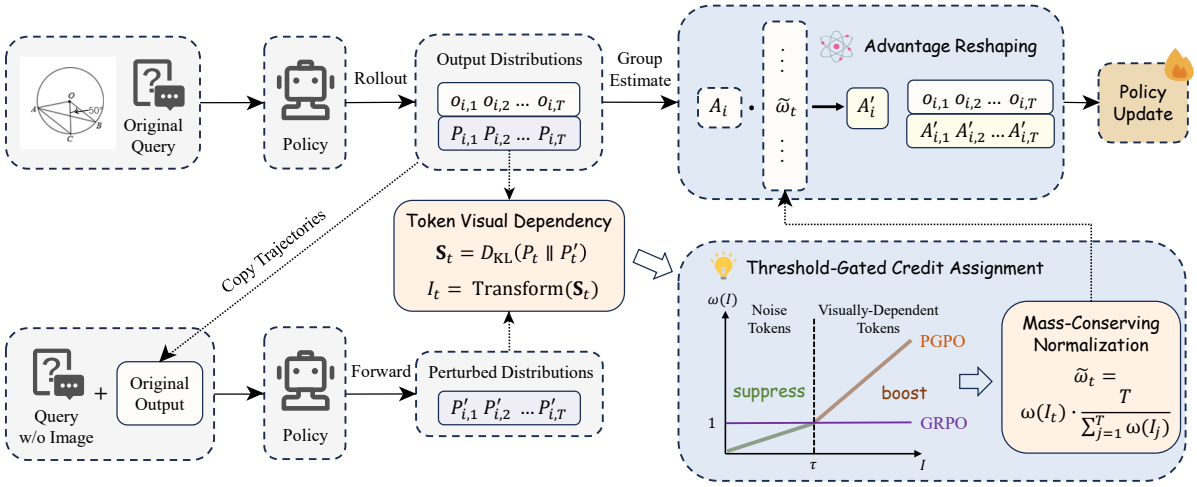


Figure 3: Overview of our proposed PGPO framework. The PGPO pipeline begins by quantifying *token visual dependency* $\mathcal{S}_t$ via KL divergence to isolate the causal information gain from visual inputs. These raw dependency signals are then transformed into bounded *token visual dependency score* I_t through logarithmic compression and min-max normalization. Finally, a threshold-gated mechanism dynamically reshapes the sequence-level GRPO advantage—amplifying learning signals for visually-dependent tokens and suppressing modality-independent noise—while applying a sum-preserving normalization to guarantee stable policy optimization.

sequence-wise min-max normalization:

$$I_t = \frac{\tilde{\mathcal{S}}_t - \min_i \tilde{\mathcal{S}}_i}{\max_i \tilde{\mathcal{S}}_i - \min_i \tilde{\mathcal{S}}_i + \epsilon} \tag{5}$$

where i indexes the generated sequence, and ϵ is a small constant preventing division by zero. This transformation resolves numerical instability and ensures invariance to baseline fluctuations across diverse multimodal prompts.

4.3 Threshold-Gated Advantage Reshaping

The execution of a fine multimodal reasoning trajectory involves a interplay between modality-independent logical derivations (yielding low I_t) and perception-grounded visual extraction (yielding high I_t). Despite this inherent token-level heterogeneity, standard GRPO uniformly broadcasts a single sequence-level advantage across these diverse tokens, which inevitably dilutes the learning signal for visually grounded reasoning.

Therefore, we propose a threshold-gated advantage reshaping mechanism. By introducing the structural prior, we formulate a piece-wise base scaling weight $\omega(I_t)$ that reallocates credit based on each token's visual dependency score:

$$\omega(I_t) = \begin{cases} \dfrac{I_t}{\tau + \epsilon}, & \text{if } I_t < \tau \\ 1 + \beta \cdot \dfrac{I_t - \tau}{1 - \tau + \epsilon}, & \text{if } I_t \geq \tau \end{cases} \tag{6}$$

where $\tau \in (0, 1)$ is a relative threshold distinguishing visually-dependent tokens from noise tokens, $\beta \geq 0$ is a boosting coefficient governing the signal amplification regime, and ϵ is a small constant ensuring numerical stability.

However, directly modifying the advantage with unbounded scaling weights risks policy update magnitudes, which can destabilize the reinforcement learning optimization. To maintain a consistent update scale, we apply a sum-preserving normalization to the base weights:

$$\tilde{\omega}_t = \omega(I_t) \cdot \frac{|o_i|}{\sum_{j=1}^{|o_i|} \omega(I_j)} \tag{7}$$

where $|o_i|$ denotes the length of the generated response. This normalization mathematically guarantees that the total advantage mass is conserved across the entire trajectory (i.e., $\sum_{t=1}^{|o_i|} \tilde{\omega}_t = |o_i|$).

For the t-th token in the i-th trajectory, the PGPO advantage is defined as:

$$\tilde{A}_{i,t} = A_i \cdot \tilde{\omega}_t \tag{8}$$

By substituting $\tilde{A}_{i,t}$ into the GRPO surrogate objective, we derive the final PGPO policy gradient formulation (omitting the clip operation):

$$\nabla_\theta \mathcal{L}_{PGPO}(\theta) = \mathbb{E}\left[\rho_{i,t}(\theta)\tilde{A}_{i,t}\nabla_\theta \log \pi_\theta(o_{i,t}|s_{i,t})\right] \tag{9}$$

By acting as an active perceptual filter, PGPO amplifies learning signals for visually-dependent tokens while dampening updates for noise tokens. This reduces gradient variance and guides the model's capacity toward mastering genuine multimodal reasoning rather than fitting linguistic prior.

4.4 Theoretical Analysis

In this section, we summarize the theoretical properties of our proposed Perception-Grounded Policy Optimization (PGPO) framework, with detailed proofs provided in the Appendix.

First, our method strictly mitigates the policy gradient variance inherent in GRPO (Appendix F). By scaling the gradient contribution of visually-independent tokens by $\varepsilon^2 \ll 1$, it acts as a causal filter that isolates valid reasoning signals and suppresses noise.

Second, we prove the necessity of advantage mass conservation ($\sum_{t=1}^{|o_i|} \tilde{\omega}_t = |o_i|$) for training stability (Appendix G). Unnormalized scaling introduces a non-zero mean shift μ, possibly triggering an $\mathcal{O}(\mu^2)$ explosion in total gradient variance.

Finally, the mapping from the raw visual dependency score $\mathcal{S}_t$ to the modulated weight $\tilde{\omega}_t$ is strictly monotonically non-decreasing (Appendix H). Furthermore, it preserves the tokens' relative ordinal importance (Appendix I), ensuring credit assignment faithfully reflects visual dependency.

5 Experiments

5.1 Setups

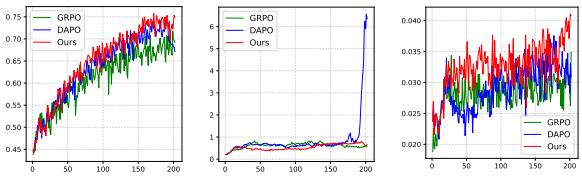
Models, Data, and Baselines. Consistent with Huang et al. (2026), we instantiate PGPO on Qwen2.5-VL-3B/7B and train with `ViRL39K` (Wang et al., 2025a), a heterogeneous dataset of multimodal reasoning tasks. For comparison, we evaluate against representative open-source reasoning LVLMs, including MM-Eureka-7B (Meng et al., 2025), VL-Rethinker-7B (Wang et al., 2025a), R1-ShareVL-7B (Yao et al., 2025), and NoisyRollout-7B (Liu et al., 2025). We also reproduce strong multimodal RLVR baselines for GRPO, DAPO, PAPO (Wang et al., 2025d), and VPPO (Huang et al., 2026).

Training Details. Following the setup of Huang et al. (2026), we train for `2` epochs and a rollout batch size of `384`. The maximum generation length is set to `2048`, and the reference KL penalty is removed. Our PGPO implementation adopts the DAPO training recipe, including dynamic sampling, clip-higher, and token-level policy-gradient optimization. The bi-level threshold and boosting hyperparameters are fixed to $\tau = 0.4$ and $\beta = 2.0$, respectively. Additional implementation details are provided in Appendix A.

Evaluation Benchmarks. We perform a broad evaluation across seven multimodal reasoning benchmarks. As in Wang et al. (2025d), we use exact-match scoring throughout, avoiding dependence on LLM-as-a-judge evaluation. These benchmarks cover math, geometry, logic, and multidisciplinary reasoning: DynaMath (Zou et al., 2024), Geo3k (Lu et al., 2021), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024a), MMK12 (Meng et al., 2025), LogicVista (Xiao et al., 2024), and MMMU-Pro (Yue et al., 2024) (detailed statistics are given in Appendix K). We report average accuracy@8 with inference temperature fixed at 1.0.

5.2 Main Results

Strong Performance. As detailed in Table 1, PGPO consistently establishes a new state-of-the-art among existing methods. Scaling effectively across both 3B and 7B parameter, our token-level advantage reshaping approach outperforms the next-best baseline, VPPO. Crucially, these performance gains are most pronounced on benchmarks demanding rigorous spatial reasoning and high visual dependency, such as Geo3k, MathVerse, and MMMU-Pro. While standard sequence-level RL methodologies struggle to assign proper credit during complex, vision-centric problem solving, our method ensures that pivotal visual anchors are accurately rewarded, translating to superior multi-step reasoning capabilities on highly multimodal tasks.



(a) Accuracy. (b) Entorpy. (c) Avg $\tilde{\mathcal{S}}$.

Figure 4: Training dynamics on Qwen2.5-VL-7B.

Training Stability. The effectiveness of PGPO is underpinned by superior training dynamics, as illustrated in the training curves against the baselines (Figure 4a), which demonstrates that PGPO

Model	Geo3k	MMK12	MathVerse	DynaMath	MathVision	LogicVista	MMMU-Pro	MathVerse_V	Avg.
MM-Eureka-7B	40.30	67.68	67.15	55.01	27.13	46.30	30.30	62.41	49.54
VL-Rethinker-7B	40.70	68.30	68.80	54.97	27.91	46.47	37.13	64.96	51.16
NoisyRollout-7B	**51.98***	50.00	67.80	54.89	22.10	47.49	34.50	63.78	49.07
R1-ShareVL-7B	41.20	70.94	68.15	54.80	26.13	45.76	35.10	64.34	50.80
Qwen2.5-VL-7B	37.29	43.04	38.36	48.15	17.58	43.01	26.67	33.63	35.97
+ GRPO	30.91	76.45	68.88	54.84	27.66	46.81	36.84	64.78	50.90
+ DAPO	41.47	78.74	67.95	55.96	28.16	47.85	36.99	64.38	52.69
+ PAPO	43.97	80.58	69.01	56.19	28.13	47.46	38.38	64.92	53.58
+ VPPO	44.38	80.11	70.95	57.11	28.27	47.52	38.75	65.54	54.08
+ **PGPO**	45.20	**80.83**	**71.45**	**57.71**	**29.02**	**47.93**	**39.01**	**66.41**	**54.70**
Qwen2.5-VL-3B	19.65	37.37	33.16	32.88	13.88	30.17	20.95	30.31	27.30
+ GRPO	32.19	62.95	58.39	47.82	25.59	40.85	28.02	54.36	43.77
+ DAPO	33.80	63.36	58.76	45.86	26.19	40.63	27.20	54.89	43.84
+ PAPO	35.11	63.54	60.81	47.30	26.37	42.70	28.76	56.90	45.19
+ VPPO	35.36	63.76	61.34	47.75	26.40	42.03	28.83	57.02	45.31
+ **PGPO**	**36.11**	**64.74**	**62.06**	**48.45**	**26.90**	**43.20**	**29.33**	**57.24**	**46.00**

Table 1: **Main Results (avg@8 acc %).** MathVerse_V refers to the vision-centric subset of MathVerse. *NoisyRollout is trained using the training set of Geo3k.

exhibits faster initial convergence, achieving higher performance more efficiently. Beyond this, another advantage of PGPO lies in its inherent optimization stability. As observed in Figure 4, standard approaches like DAPO suffer from severe late-stage training collapse on Qwen2.5-VL-7B, characterized by a sharp drop in accuracy and a corresponding spike in policy entropy (note: DAPO results are reported from the optimal pre-collapse checkpoint to ensure a fair comparison). While contemporary baselines such as PAPO and VPPO attempt to mitigate this instability by introducing heuristic entropy penalty terms to their loss functions, PGPO addresses the underlying mathematical root cause. By executing fine-grained advantage reallocation for pivot tokens, PGPO reduces gradient variance and stabilizes the training trajectory, eliminating the need for ad-hoc entropy regularization.

Visual Dependency. Furthermore, as shown in Figure 4c, the average visual dependency ($\tilde{\mathcal{S}}_t$) of the generated tokens increases alongside the model's training accuracy throughout the RLVR process. This indicates that the model is fundamentally enhancing its visual perception to facilitate multimodal reasoning. However, the extent of this perceptual enhancement heavily depends on the underlying optimization strategy. Under our PGPO framework, visual dependency exhibits a rapid and steady upward trend, whereas baselines like DAPO stagnate during the late training phase. This divergence demonstrates that our targeted bi-level learning signal acts as a potent implicit regularizer, actively driving the policy to rely on visual evidence and enabling a highly efficient, robust path to rigorous multimodal reasoning.

6 Analysis

6.1 Effect of the Threshold-Gated Mechanism

To systematically validate the structural necessity of our proposed threshold-gated advantage reshaping mechanism and its sum-preserving normalization, we conduct an ablation study evaluating isolated variants of PGPO on Qwen2.5-VL-3B.

Variant	MathVerse	MMMU-Pro
DAPO	58.76	27.20
Suppression-Only ($\tau = 1$)	60.34	28.71
Boosting-Only ($\tau = 0$)	59.04	28.15
PGPO w/o Norm.	60.18	28.10
PGPO	**62.06**	**29.33**

Table 2: Ablation of advantage modulation strategies.

As presented in Table 2, the full PGPO framework substantially outperforms all other variants. Deconstructing the gating mechanism reveals that the suppression-only strategy achieves stronger performance than the boosting-only variant. This indicates that actively reducing the gradient of visually irrelevant syntactic tokens is more critical for

multimodal optimization than merely amplifying perceptual anchors.

Furthermore, the necessity of Equation 7 is empirically verified by the *PGPO w/o Norm.* variant. Removing the sum-preserving normalization significantly degrades performance, empirically corroborating our mathematical derivation that unnormalized advantage scaling disrupts policy update constraints via a non-zero mean shift. Ultimately, integrating threshold-gated suppression, signal boosting, and normalization into the full PGPO configuration yields the optimal synergy.

6.2 Hyperparameter Analysis

We further investigate the sensitivity of our method to its core gating hyperparameters on Qwen2.5-VL-3B: the threshold τ for balancing noise suppression and signal amplification, and the boosting coefficient β for scaling score reallocation in Table 3.

Threshold (τ)	Boosting (β)	MathVerse	MMMU-Pro
0.2	1.0	58.42	27.15
0.2	2.0	59.18	27.89
0.4	1.0	60.36	28.98
0.4	**2.0**	**62.06**	**29.33**
0.6	1.0	59.12	28.40
0.6	2.0	56.11	28.48

Table 3: Hyperparameter sensitivity analysis results.

A permissive threshold ($\tau = 0.2$) dilutes credit assignment by boosting redundant tokens. Conversely, a relatively strict threshold ($\tau = 0.6$) yields a sparse advantage signal that, when paired with high boosting ($\beta = 2.0$), destabilizes PPO clipping bounds and degrades performance. Ultimately, $\tau = 0.4$ precisely partitions the trajectory, safely isolating the $\beta = 2.0$ amplification to true perceptual anchors without inflating global gradients.

6.3 Performance on General Benchmarks

Method	MMBench	MMStar	MME
Qwen2.5-VL-3B	74.00	43.05	69.27
GRPO	83.88	57.34	75.36
DAPO	84.05	57.76	75.47
PGPO (Ours)	**84.96**	**58.43**	**77.43**

Table 4: General benchmarks evaluation on Qwen2.5-VL-3B. Performance is reported as avg@8 accuracy.

To ensure our method does not impair general visual-language capabilities, we evaluated its performance on three unseen, general VQA benchmarks: MMBench (Liu et al., 2024), MMStar (Chen et al., 2024a) and MME (Fu et al., 2025).

As shown in Table 4, PGPO consistently outperforms baseline methods, demonstrating a general superiority in vision tasks requiring fine-grained perception or reasoning. This confirms that anchoring advantage redistribution to token-level visual dependencies prevents the policy from overfitting to superficial textual patterns. Meanwhile, PGPO inherently sharpens perceptual grounding without sacrificing general domain knowledge.

6.4 Analysis of Computational Overhead

To assess the extra computation introduced by the second forward pass for token visual dependency estimation, we measure training-time overhead relative to the DAPO baseline. Table 5 shows that the added cost is limited to about 10%. This efficiency comes from computing token probabilities for the full sequence in one parallel pass.

In addition, we run DAPO under the same time budget (denoted as DAPO_E). As reported in Table 5, DAPO_E shows little improvement despite longer training, while PGPO delivers a 2-point average gain. These results indicate that token-level shaping with visual dependency signals improves learning effectiveness on complex visual reasoning, making PGPO's small computational overhead a worthwhile trade-off for performance gains.

Method	Training Hours / Relative Cost	Avg. Performance
DAPO	25.4 / 1.00×	43.84
DAPO_E	28.1 / 1.11×	43.93
PGPO	28.1 / 1.11×	**46.00**

Table 5: Analysis results of computational overhead on Qwen2.5-VL-3B with 2 NVIDIA H100 80G GPUs.

7 Conclusion

In this work, we propose Perception-Grounded Policy Optimization (PGPO), which is dynamically reshapes token-level advantages through a threshold-gated, sum-preserving mechanism, actively amplifying learning signals for perceptual anchors while suppressing modality-independent gradient noise. Extensive evaluations across seven complex multimodal benchmarks demonstrate the effectiveness of PGPO. Both theoretical and empirical analyses confirm that PGPO significantly reduces gradient variance and yields highly robust, perception-grounded multimodal reasoning.

Limitations

While PGPO significantly advances multimodal RLVR, we acknowledge certain limitations for future research. While the hyperparameters of our threshold-gated mechanism (threshold τ and boosting coefficient β) generalize robustly across our benchmarks, they were optimized for our specific training distribution and may require re-tuning when applied to other datasets or model architectures. Furthermore, limited by our computational resources, our experiments validate PGPO on models up to the 7B parameter scale. Although these results indicate a positive scaling trend, its efficacy on large-scale models remains to be verified.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. 2020. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Preprint*, arXiv:1908.00261.

Dilip Arumugam, Peter Henderson, and Pierre-Luc Bacon. 2021. An information-theoretic perspective on credit assignment in reinforcement learning. *Preprint*, arXiv:2103.06224.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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

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

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

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

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

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. 2025c. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024b. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *Preprint*, arXiv:2403.00425.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. 2026. Spotlight on token perception for multimodal reinforcement learning. *Preprint*, arXiv:2510.09285.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306.

Sham M. Kakade. 2001. A natural policy gradient. In *Neural Information Processing Systems*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

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

Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, Bo Zheng, and Junchi Yan.

2025b. Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *Preprint*, arXiv:2510.13554.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

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

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

John Schulman. 2020. Approximating KL divergence. Blog post.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Richard S. Sutton and Andrew G. Barto. 1998. Reinforcement learning: An introduction. *IEEE Trans. Neural Networks*, 9:1054–1054.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. 2025b. Emergent hierarchical reasoning in llms through reinforcement learning. *Preprint*, arXiv:2509.03646.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, and 1 others. 2025c. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and 1 others. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

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

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. 2025a. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *Preprint*, arXiv:2505.11821.

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, and 1 others. 2025b. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *arXiv preprint arXiv:2507.05255*.

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

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

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

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and 1 others. 2025. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, and 1 others. 2025. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*.

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

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

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. 2025. Easyr1: An efficient, scalable, multi-modality rl training framework.

Meng Zhou, Ziyu Liu, Pengwei Sui, Yixuan Li, and Yuk Ying Chung. 2020. Learning implicit credit assignment for cooperative multi-agent reinforcement learning. *Preprint*, arXiv:2007.02529.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*.

Appendix

A Implementation Details 12

B PGPO on the GRPO algorithm 13

C PGPO on Qwen3-VL 13

D Detailed Settings for Empirical Analysis of Visual Dependency 13

E Information-Theoretic Foundation of the Visual Dependency Metric 15

F Theoretical Analysis of Gradient Noise Suppression 16

G Theoretical Justification for Sum-Preserving Normalization 17

H Monotonicity Analysis of $\tilde{\omega}_t$ 19

I Rank-Preserving Property of $\tilde{\omega}_t$ 20

J Analysis of the Training Dataset 21

K Analysis of Evaluation Benchmarks 21

L LLM Usage Statement 21

M Case Study 22

A Implementation Details

Training Details. We implement our method on top of the EasyR1 framework (Zheng et al., 2025; Sheng et al., 2024), and run all experiments with PyTorch 2.8.0 and CUDA 12.6. Our base models are the open-source Qwen2.5-VL-3B and Qwen2.5-VL-7B checkpoints. Each model is trained for two epochs on `ViRL39K` (Wang et al., 2025a), with the vision tower kept trainable throughout optimization. During online RL, we sample 5 responses per prompt. The final reward is a weighted sum of accuracy and format compliance, defined as $R = 0.9S_{\text{acc}} + 0.1S_{\text{format}}$, where $S_{\text{acc}} \in \{0, 1\}$ indicates answer correctness and $S_{\text{format}} \in \{0, 1\}$ indicates whether the output satisfies the required response format. Our objective follows the DAPO training recipe, including dynamic sampling, clip-higher, and token-level policy-gradient optimization, without applying KL regularization. Under a unified setup, we also reproduce PAPO and VPPO for direct comparison. In the main experiments, all runs use the same bi-level threshold $\tau = 0.4$ and boosting coefficient $\beta = 2.0$ for both training and evaluation. For fair comparison on Qwen2.5-VL-7B, we report DAPO results from the best pre-collapse checkpoint (step 190 out of 202), since entropy explosion appears near the end of training. All experiments are conducted on NVIDIA H100 80GB GPU clusters. Complete optimizer, RL, and evaluation hyperparameters are listed in Table 6.

Prompt Template. Across both training and evaluation, we adopt the standardized prompt template shown below. This format is designed to induce stable Chain-of-Thought reasoning and to support reliable automatic extraction 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{}`.

Table 6: Configuration summary for model optimization and inference.

Category	Setting	Category	Setting
Optimizer	AdamW	Global batch	128
Learning rate	1×10^{-6}	Rollout batch	384
LR policy	Fixed	Samples per prompt	5
Training epochs	2	Rollout top-p	0.99
Vision encoder	Trainable	Max response length	2048
Sampling strategy	Dynamic	Inference temperature	1.0
Lower clip bound	0.2	Inference top-p	1.0
Upper clip bound	0.28	Max new tokens	2048
Loss reduction	Token-level	Filtering criterion	Overall
KL regularizer	Not used	Reward composition	$0.9S_{\text{acc}} + 0.1S_{\text{format}}$
PGPO threshold τ	0.4	Boosting factor β	2.0

A.1 PGPO Details

In Section 3, token visual dependency $\mathcal{S}(s_t, I)$ is theoretically defined as the Kullback-Leibler (KL) divergence over the entire vocabulary space $\mathcal{V}$. However, computing the exact full-vocabulary KL divergence in Large Vision-Language Models (LVLMs) is practically prohibitive. To ensure high training efficiency and stable optimization, we introduce two critical engineering optimizations: attention masking for the unconditioned forward pass, and a low-variance, unbiased Monte-Carlo estimator for the KL divergence.

Visually-Unconditioned Distribution via Attention Masking. To obtain the unconditioned probability distribution $\pi_\theta(\cdot|s_t)$, one naive approach is to replace the image with a zero-tensor (blank image) or remove the visual tokens from the input sequence entirely. However, altering the input sequence length fundamentally shifts the absolute positional encodings of the subsequent text tokens, introducing confounding variables that distort the pure measurement of visual dependency.

Instead, we compute the unconditioned distribution by manipulating the causal attention mask during a secondary forward pass. Specifically, we set the attention mask corresponding to all visual tokens to `False`, effectively blinding the text tokens to the visual prefix while preserving their original positional indices. It avoids the overhead of re-tokenization and KV-cache structural modifications, allowing the unconditioned logits to be computed seamlessly using the exact same tensor shapes as the primary forward pass.

Estimation of KL Divergence. Mathematically, the exact KL divergence at step t requires computing a sum over the entire vocabulary $\mathcal{V}$ (typically $> 100,000$ tokens):

$$D_{\mathrm{KL}}(\pi_I \parallel \pi_\emptyset) = \sum_{x \in \mathcal{V}} \pi_I(x) \log \frac{\pi_I(x)}{\pi_\emptyset(x)} \qquad (10)$$

where, for brevity, we denote the visually-conditioned policy as $\pi_I(x) := \pi_\theta(x|s_t, I)$ and the unconditioned policy as $\pi_\emptyset(x) := \pi_\theta(x|s_t)$. In practice, explicitly computing this full sum incurs massive GPU memory footprint and I/O overhead. Furthermore, it aggregates the numerical noise from the long tail of highly improbable tokens, which can destabilize the dependency signal.

To resolve this, following Schulman (2020), we approximate the expectation using Monte-Carlo sampling based solely on the generated token $x_t \sim \pi_I$. Let the probability ratio be defined as $r = \frac{\pi_\emptyset(x_t)}{\pi_I(x_t)}$. We use $(r-1) - \log r$ as an unbiased, strictly non-negative, and low-variance estimator for KL divergence.

B PGPO on the GRPO algorithm

To evaluate the generalizability of PGPO, we integrated it with GRPO. As shown in Table 7, PGPO also improves GRPO's accuracy, surpassing DAPO. This gain aligns with the improvement observed when applying PGPO to DAPO (Table 1), confirming that the performance enhancements are intrinsic to our visually-perceptive optimization strategy rather than artifacts of the underlying base policy gradient algorithm.

C PGPO on Qwen3-VL

To examine whether PGPO generalizes beyond the Qwen2.5-VL family, we further evaluate it on Qwen3-VL-2B-Instruct (Bai et al., 2025a). As shown in Table 8, PGPO consistently improves DAPO across all benchmarks, increasing the average accuracy from 48.18 to 49.23. Our experimental setup is identical to that of the main experiment, which simultaneously demonstrates the generalizability of our hyperparameters. The gains span mathematical, logical, and discipline-intensive reasoning tasks, indicating that token-level visual dependency remains an effective credit-assignment signal under a different LVLM backbone.

D Detailed Settings for Empirical Analysis of Visual Dependency

In Section 3.2, we presented empirical analyses demonstrating the distributional sparsity of token visual dependency ($\mathcal{S}$), its strong correlation with semantic visual anchors, and its inverse relationship with multimodal hallucinations. To ensure full reproducibility and provide deeper insights into our methodology, this section details the experimental setups, data processing pipelines, and evaluation metrics used for these preliminary analyses. All empirical analyses were conducted using the Qwen2.5-VL-3B model. To ensure deterministic and reproducible trajectories, we employed greedy decoding for all text generation tasks.

D.1 Distributional Sparsity and Content Token Analysis

Dataset Selection. We utilized the vision-dominant subset of the MathVerse benchmark. This specific subset comprises problems where the visual information is mathematically indispensable, ensuring that a reliance on language priors alone cannot yield the correct answer. We processed the multi-choice test-mini split.

Part-of-Speech (POS) Categorization. To analyze the discrepancy in visual dependency between different types of tokens, we employed the `spaCy` to perform POS tagging on the generated trajectories. *Content Tokens* were defined as tokens categorized under structural and semantic classes

Model	Geo3k	MMK12	MathVerse	DynaMath	MathVision	LogicVista	MMMU-Pro	MathVerse$_V$	Avg.
Qwen2.5-VL-3B	19.65	37.37	33.16	32.88	13.88	30.17	20.95	30.31	27.30
+ GRPO	32.19	62.95	58.39	47.82	25.59	40.85	28.02	54.36	43.77
+ DAPO	33.80	63.36	58.76	45.86	**26.19**	40.63	27.20	54.89	43.84
+ PGPO w/ GRPO	**33.81**	**63.96**	**59.62**	**47.91**	26.00	**41.53**	**28.29**	**56.75**	**44.73**

Table 7: Results on PGPO by applying it to GRPO (avg@8 acc %).

Model	Geo3k	MMK12	MathVerse	DynaMath	MathVision	LogicVista	MMMU-Pro	MathVerse$_V$	Avg.
Qwen3-VL-2B	28.10	22.95	31.13	38.65	8.39	27.43	20.56	29.17	25.80
+ GRPO	43.88	53.16	63.93	55.33	28.68	45.41	30.39	59.92	47.59
+ DAPO	45.26	52.29	64.91	55.76	28.61	46.76	31.15	60.73	48.18
+ PGPO	**46.73**	**53.81**	**66.04**	**56.32**	**29.15**	**48.19**	**31.94**	**61.68**	**49.23**

Table 8: Results of PGPO on Qwen3-VL (avg@8 acc %).

critical to reasoning: 'NUM' (numerals), 'NOUN' (nouns/geometric concepts), 'VERB' (verbs), and 'ADJ' (adjectives). *Other Tokens* are other remaining tokens. Tokens split into sub-words by the tokenizer inherited the POS tag of their parent word. As shown in Figure 2(b), we computed the mean $\mathcal{S}$ for all tokens within these respective macro-categories across all trajectories. As shown in Figure 5, tokens with high visual dependence often carry critical visual semantics.

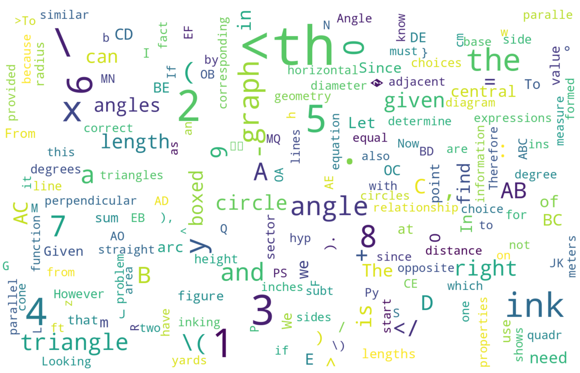


Figure 5: Top 200 $\mathcal{S}$ tokens word cloud on all generated tokens of vision-dominant MathVerse.

D.2 Visual Anchor Annotation

To evaluate whether $\mathcal{S}$ is associated with semantic visual grounding, we constructed an annotation pipeline to identify "Visual Anchors"—tokens highly requiring image observation for inference.

Annotation. We used complete trajectories from the vision-dominant MathVerse generation set. Given the complexity of mathematical reasoning, we leveraged GPT-5 as a zero-shot annotator. GPT-5 was provided with the original image, the textual question, and the generated trajectory, and was prompted to extract specific words that act as visual anchors. Prompt for LLM annotation is: *You are an expert annotator for multimodal mathematical reasoning. Given the following Image, Question, and the Model's Generated Response, your task is to identify 'Visual Anchors' within the response. A Visual Anchor is defined as a specific entity, number, spatial relationship, or geometric attribute mentioned in the text that CANNOT be properly deduced from the text alone and strictly requires extracting information from the image. Output your response as a strictly formatted JSON list of words/phrases extracted directly from the generation.*

Token Mapping. The extracted anchor phrases were string-matched back to the generated token sequence. Tokens comprising these phrases were labeled as *Visual Anchor tokens*, while all remaining tokens were labeled as *Other tokens*. The statistical significance of the $\mathcal{S}$ differential between these two groups also was confirmed via a two-sample t-test ($p < 0.001$), validating our core hypothesis that high $\mathcal{S}$ corresponds to critical visual grounding.

D.3 Hallucination Analysis on the CHAIR Benchmark

To understand the relationship between visual dependency and factual correctness, we extended our analysis to the standard object hallucination benchmark, CHAIR (Caption Hallucination Assessment with Image Relevance) (Rohrbach et al., 2019).

We sampled 5,000 images from the MSCOCO

(Lin et al., 2015) 2014 validation set. For each image, we prompted Qwen2.5-VL-3B with: "Describe this image in detail." We employed the official CHAIR evaluation script, leveraging MSCOCO ground-truth object annotations (synsets), to assess the generated sequences. First, we parsed the generated caption to extract all mentioned objects using the standard CHAIR synonym dictionary. Subsequently, these extracted objects are cross-referenced against the image's ground-truth annotations; tokens corresponding to objects present in the ground truth are classified as truthful, whereas those representing absent objects are designated as hallucinated.

We isolated these two subsets of tokens across the 5,000 captions and computed their average $\mathcal{S}$ values. The finding (Figure 2(d)) that hallucinated tokens exhibit a significantly lower visual dependency provides evidence that multimodal hallucinations in LVLMs are largely driven by over-reliance on language priors rather than visual processing. This conclusion is consistent with existing research findings (Leng et al., 2023; Chen et al., 2024b).

E Information-Theoretic Foundation of the Visual Dependency Metric

In this appendix, we establish the theoretical soundness of using the Kullback-Leibler (KL) divergence to quantify visual dependency, denoted as $\mathcal{S}(s_t, I)$. Specifically, we demonstrate that the proposed formulation in Section 3.1 is mathematically equivalent to the state-specific conditional information gain provided by the visual input.

E.1 Conditional Mutual Information in Autoregressive Generation

In information theory, the reduction in uncertainty about a random variable X given the observation of another random variable Y is measured by Mutual Information $I(X;Y)$. When a prior context Z is already known, the additional information provided by Y is captured by the Conditional Mutual Information:

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z), \quad (11)$$

where $H(\cdot)$ denotes Shannon entropy.

Within the context of a Large Vision-Language Model (LVLM), we define the following random variables:

- $O_t \in \mathcal{V}$: The next token to be generated from the vocabulary $\mathcal{V}$.
- $S_t = (q, o_{<t})$: The current state, comprising the textual query and past generated tokens.
- $\mathcal{I}$: The visual input provided to the model.

Under the distribution defined by the model policy π_θ, the expected information gain regarding the next token O_t provided by the image $\mathcal{I}$, conditioned on the text context S_t, is strictly given by $I_{\pi_\theta}(O_t; \mathcal{I} \mid S_t)$.

E.2 Derivation of Visual Dependency via KL Divergence

By expanding the definition of Conditional Mutual Information (Equation 11) using conditional entropy, we can rewrite $I_{\pi_\theta}(O_t; \mathcal{I} \mid S_t)$ in terms of KL divergence.

Let the visually-conditioned policy be $\pi_\theta(\cdot \mid s_t, I)$ and the visually-marginalized (text-only) policy be $\pi_\theta(\cdot \mid s_t)$. The conditional mutual information over the joint distribution of states and images can be expressed as:

$$\begin{aligned} &I_{\pi_\theta}(O_t; \mathcal{I} \mid S_t) = \\ &\mathbb{E}_{s_t, I \sim \pi_\theta(S_t, \mathcal{I})} \left[\sum_{o_t \in \mathcal{V}} \pi_\theta(o_t \mid s_t, I) \log \frac{\pi_\theta(o_t \mid s_t, I)}{\pi_\theta(o_t \mid s_t)} \right] \\ &= \mathbb{E}_{s_t, I \sim \pi_\theta(S_t, \mathcal{I})} \left[D_{\mathrm{KL}} \Big(\pi_\theta(\cdot \mid s_t, I) \parallel \pi_\theta(\cdot \mid s_t) \Big) \right]. \end{aligned} \quad (12)$$

Equation 12 demonstrates that the global information gain—averaged across all possible reasoning states and images—is exactly the expected KL divergence between the visually-conditioned policy and the text-only policy.

During RLVR inference, we evaluate the specific informational value of a *given* image instance I at a *given* reasoning state s_t, rather than the expectation over the entire distribution. Removing the outer expectation from Equation 12 yields the state-specific conditional information gain:

$$\begin{aligned} \mathcal{S}(s_t, I) &:= I_{\pi_\theta}(O_t; \mathcal{I} = I \mid S_t = s_t) \\ &= D_{\mathrm{KL}} \Big(\pi_\theta(\cdot \mid s_t, I) \parallel \pi_\theta(\cdot \mid s_t) \Big). \end{aligned} \quad (13)$$

This confirms that our visual dependency metric $\mathcal{S}(s_t, I)$ fundamentally represents the exact bits of information contributed causally by the image to the prediction of the t-th token.

E.3 Superiority of KL Divergence

A seemingly intuitive alternative to quantify visual dependency is to compute the direct reduction

in entropy before and after observing the image: $\Delta H = H(\pi_\theta(\cdot \mid s_t)) - H(\pi_\theta(\cdot \mid s_t, I))$. However, we exclusively employ KL divergence due to two fundamental theoretical advantages:

Capturing Distributional Shift. Entropy purely measures the *volume* of uncertainty, ignoring the semantic *direction* of the probability shift. Consider a scenario where the text-only policy is highly confident but factually incorrect (e.g., hallucinating a common object based on language priors), yielding a very low entropy $H(\pi_\theta(\cdot \mid s_t))$. When the image is introduced, the model correctly grounds its prediction on visual evidence, yielding a different token but with similarly high confidence (low $H(\pi_\theta(\cdot \mid s_t, I))$). In this case, the entropy difference is near zero ($\Delta H \approx 0$), failing to detect the critical intervention of the visual modality. In contrast, KL divergence measures the cross-entropy penalty; because the original text prior assigned low probability to the visually-grounded truth, the KL divergence will be heavily magnified, perfectly capturing the visual correction.

Strict Non-Negativity for Stable Credit Assignment. Depending on the clarity of the image, the visual input might actually introduce ambiguity compared to an overconfident language prior, which makes ΔH negative. Using a metric that can be negative severely complicates advantage reshaping and introduces instability in RL credit assignment. By Gibbs' inequality, KL divergence mathematically guarantees $\mathcal{S}(s_t, I) \geq 0$. This ensures a strict, well-behaved lower bound where 0 strictly corresponds to identical distributions (zero visual reliance).

F Theoretical Analysis of Gradient Noise Suppression

In vanilla GRPO, the same trajectory-level scalar advantage is broadcast to every decoding position. This design is statistically inefficient when many positions are weakly grounded in the image: those tokens still receive full-magnitude updates even though their contribution to the visual task objective is limited. Here we show that PGPO attenuates this irrelevant gradient component and provably reduces the corresponding noise term, which is adapted from Huang et al. (2026).

Problem Setup and Assumptions. Consider one sampled response $y_{1:T} = (y_1, \ldots, y_T)$ under multimodal context $c = (I, q)$. Define the token score vector

$$\mathbf{u}_t := \nabla_\theta \log \pi_\theta(y_t \mid c, y_{<t}). \tag{14}$$

Using the dependency indicator I_t, we split indices into two subsets:

$$\mathcal{V} := \{t \mid I_t \geq \tau\}, \qquad \mathcal{B} := \{t \mid I_t < \tau\}, \tag{15}$$

where $\mathcal{V}$ denotes visually grounded positions and $\mathcal{B}$ denotes background (nuisance) positions.

To characterize estimator noise, we adopt the following assumptions:

Assumption 1 (Cross-Time Decorrelation): For $t \neq j$, $\mathbb{E}[\mathbf{u}_t^\top \mathbf{u}_j] \approx 0$.

Assumption 2 (Scalar-Gradient Independence): The trajectory advantage A is independent of $\{\mathbf{u}_t\}_{t=1}^T$.

Assumption 3 (Weight-Gradient Weak Coupling): The PGPO weight $\tilde{\omega}_t$ is approximately independent of $\mathbf{u}_t$.

Assumption 4 (Second-Moment Proxy): In high dimension, $\mathrm{Var}(\mathbf{g})$ is well-approximated by the magnitude of $\mathbb{E}[\|\mathbf{g}\|_2^2]$.

Theorem F.1 (PGPO Noise Suppression)**.** *Under Assumptions 1–4, if PGPO enforces $\tilde{\omega}_k \leq \varepsilon$ on $k \in \mathcal{B}$ with $0 < \varepsilon \ll 1$, then the nuisance-token component in the gradient second moment is reduced by at least a multiplicative factor ε^2 relative to GRPO.*

Proof. We compare the squared-norm second moments of the two estimators.

1. GRPO baseline. The trajectory gradient in GRPO is

$$\mathbf{g}_{\text{grpo}} = A \sum_{t=1}^{T} \mathbf{u}_t. \tag{16}$$

$$
\begin{aligned}
\mathbb{E}\big[\|\mathbf{g}_{\text{grpo}}\|_2^2\big] &= \mathbb{E}\left[\left\|A\sum_{t=1}^{T}\mathbf{u}_t\right\|_2^2\right] \\
&\overset{\text{Asm. 2}}{=} \mathbb{E}[A^2]\,\mathbb{E}\left[\left\|\sum_{t=1}^{T}\mathbf{u}_t\right\|_2^2\right] \\
&= \mathbb{E}[A^2]\,\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{u}_t\|_2^2 + \sum_{t\neq j}\mathbf{u}_t^\top\mathbf{u}_j\right] \\
&\overset{\text{Asm. 1}}{\approx} \mathbb{E}[A^2]\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{u}_t\|_2^2] \\
&= \mathbb{E}[A^2]\left(\sum_{t\in\mathcal{V}}\mathbb{E}[\|\mathbf{u}_t\|_2^2] + \sum_{k\in\mathcal{B}}\mathbb{E}[\|\mathbf{u}_k\|_2^2]\right).
\end{aligned}
\tag{17}
$$

The second summation in Eq. 17 is pure nuisance contribution: it is nonzero in expectation even when those tokens are weakly related to visual reasoning.

2. PGPO estimator. PGPO applies token-wise modulation:

$$
\mathbf{g}_{\text{pgpo}} = A\sum_{t=1}^{T}\tilde{\omega}_t\mathbf{u}_t. \tag{18}
$$

Then

$$
\begin{aligned}
\mathbb{E}\big[\|\mathbf{g}_{\text{pgpo}}\|_2^2\big] &= \mathbb{E}\left[\left\|A\sum_{t=1}^{T}\tilde{\omega}_t\mathbf{u}_t\right\|_2^2\right] \\
&\overset{\text{Asm. 2}}{=} \mathbb{E}[A^2]\,\mathbb{E}\left[\sum_{t=1}^{T}\|\tilde{\omega}_t\mathbf{u}_t\|_2^2 + \sum_{t\neq j}\tilde{\omega}_t\tilde{\omega}_j\mathbf{u}_t^\top\mathbf{u}_j\right] \\
&\overset{\text{Asm. 1}}{\approx} \mathbb{E}[A^2]\sum_{t=1}^{T}\mathbb{E}[\tilde{\omega}_t^2\|\mathbf{u}_t\|_2^2] \\
&\overset{\text{Asm. 3}}{=} \mathbb{E}[A^2]\sum_{t=1}^{T}\mathbb{E}[\tilde{\omega}_t^2]\,\mathbb{E}[\|\mathbf{u}_t\|_2^2] \\
&= \mathbb{E}[A^2] \\
&\left(\sum_{t\in\mathcal{V}}\mathbb{E}[\tilde{\omega}_t^2]\mathbb{E}[\|\mathbf{u}_t\|_2^2] + \sum_{k\in\mathcal{B}}\mathbb{E}[\tilde{\omega}_k^2]\mathbb{E}[\|\mathbf{u}_k\|_2^2]\right).
\end{aligned}
\tag{19}
$$

3. Noise bound and interpretation. By construction, PGPO keeps $\tilde{\omega}_t$ around unit scale on $t \in \mathcal{V}$ and enforces $\tilde{\omega}_k \leq \varepsilon$ on $k \in \mathcal{B}$. Plugging this into Eq. 19 gives

$$
\begin{aligned}
&\mathbb{E}\big[\|\mathbf{g}_{\text{pgpo}}\|_2^2\big] \leq \\
&\mathbb{E}[A^2]\left(\sum_{t\in\mathcal{V}}\mathbb{E}[\|\mathbf{u}_t\|_2^2] + \varepsilon^2\sum_{k\in\mathcal{B}}\mathbb{E}[\|\mathbf{u}_k\|_2^2]\right).
\end{aligned}
\tag{20}
$$

Combined with Assumption 4, the bound shows a clean separation: the useful visually grounded component is maintained, while the nuisance component is quadratically damped by ε^2 compared with Eq. 17. Therefore, in the small-ε rule, PGPO behaves as a perception-aware variance filter that suppresses updates driven by modality-irrelevant tokens. □

G Theoretical Justification for Sum-Preserving Normalization

In the PGPO framework, we dynamically rescale token-level advantages based on visual dependency. However, directly applying unnormalized weights risks fundamentally destabilizing the training process. In this section, we first explain the structural flaw of unnormalized advantage scaling, thereby motivating our sum-preserving normalization mechanism. We then provide a mathematical proof demonstrating the possible variance explosion that occurs if this normalization is omitted.

G.1 The Flaw of Unnormalized Scaling

Let A_i represent the standard GRPO group-normalized sequence advantage for the i-th sequence in a sampled group of size G. By design, GRPO standardizes rewards such that the advantages maintain a zero mean within the group: $\sum_{i=1}^{G} A_i = 0$. This implicit zero-sum baseline is critical for reducing the variance of the policy gradient estimator (Agarwal et al., 2020; Kakade, 2001).

Suppose we naively introduce unnormalized advantage scaling. Whether applied via a sequence-level coefficient $c_i > 0$ or via token-level weights that do not sum to the sequence length ($\sum_{t=1}^{T_i}\omega_t \neq T_i$), this effectively scales the net advantage of the trajectory, yielding a modulated advantage $\tilde{A}_i = c_i A_i$. Substituting $\tilde{A}_i$ into the standard policy gradient objective, the unclipped gradient estimator

becomes:

$$\mathbf{g}_{\text{scaled}} = \frac{1}{G}\sum_{i=1}^{G}\tilde{A}_i\sum_{t=1}^{T_i}\nabla_\theta \log \pi_\theta(o_{i,t} \mid s_{i,t}) = \frac{1}{G}\sum_{i=1}^{G}(c_i A_i)\sum_{t=1}^{T_i}\nabla_\theta \log \pi_\theta(o_{i,t} \mid s_{i,t}) \tag{21}$$

The flaw of this unnormalized scaling lies in the destruction of the zero-mean baseline. With the introduction of the sequence-dependent coefficient c_i, the modulated advantages no longer sum to zero:

$$\mu_{\text{scaled}} = \frac{1}{G}\sum_{i=1}^{G} c_i A_i \neq 0 \tag{22}$$

This non-zero mean shift ($\mu_{\text{scaled}} \neq 0$) may arbitrarily distorts the magnitude of the parameter updates, disrupting stable credit assignment.

To prevent this, PGPO employs sum-preserving normalization Equation 7. This mathematically guarantees that the total advantage mass distributed across the trajectory strictly equals $T_i \cdot A_i$, thereby preserving the sequence-level baseline $\mu = 0$ while effectively reallocating credit at the token level.

G.2 Formal Proof: Asymptotic Quadratic Variance Inflation Induced by Mean Shifts

To rigorously explain why preserving the zero-mean structure is important, we analyze the effect of adding a constant mean shift μ to an advantage-weighted policy-gradient estimator. The key result is that, although such a shift does not change the estimator's mean, it introduces an additional covariance term that grows quadratically in $|\mu|$ and eventually dominates the estimator variance.

Preliminaries and Definitions. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\tau \sim \pi_\theta$ denote a trajectory sampled from a policy π_θ with parameter $\theta \in \mathbb{R}^d$. Write $\mathbb{E}[\cdot] := \mathbb{E}_{\tau\sim\pi_\theta}[\cdot]$.

Definition 1 (Score Function). The trajectory score function is $\mathbf{g}(\tau) := \nabla_\theta \log \pi_\theta(\tau)$. Under standard regularity conditions, $\mathbb{E}[\mathbf{g}(\tau)] = \mathbf{0}$.

Definition 2 (Fisher Information Matrix). The Fisher Information Matrix (FIM) is $\mathbf{F} := \mathbb{E}\big[\mathbf{g}(\tau)\mathbf{g}(\tau)^\top\big] \in \mathbb{R}^{d\times d}$. Hence $\mathbf{F} \succeq 0$. If the policy is non-degenerate in all parameter directions, then $\mathbf{F} \succ 0$, so $\lambda_{\min}(\mathbf{F}) > 0$.

Definition 3 (Base and Shifted Estimators). Let $A^*(\tau)$ be a scalar random variable satisfying $\mathbb{E}[A^*(\tau)] = 0$. Define the base estimator $\hat{\mathbf{g}}_0 := A^*(\tau)\mathbf{g}(\tau)$, and, for a constant shift $\mu \in \mathbb{R}$, define the shifted estimator $\hat{\mathbf{g}}_\mu := (A^*(\tau) + \mu)\mathbf{g}(\tau)$.

Observe that

$$\mathbb{E}[\hat{\mathbf{g}}_\mu] = \mathbb{E}[A^*(\tau)\mathbf{g}(\tau)] + \mu\,\mathbb{E}[\mathbf{g}(\tau)] = \mathbb{E}[\hat{\mathbf{g}}_0]. \tag{23}$$

Therefore, adding a constant shift μ does not change the estimator mean, but it may change its covariance.

Assumption 1 (Finite Moments). Assume all expectations below are well-defined and finite. In particular, define

$$\mathbf{C} := \mathbb{E}\Big[A^*(\tau)\mathbf{g}(\tau)\mathbf{g}(\tau)^\top\Big], \tag{24}$$

and assume $\|\mathbf{C}\|_2 < \infty$ and $\mathrm{tr}(\mathbf{C})$ is finite.

Proposition G.1 (Covariance inflation due to a mean shift)**.** *Under Assumption 1,*

$$\mathrm{Cov}(\hat{\mathbf{g}}_\mu) = \mathrm{Cov}(\hat{\mathbf{g}}_0) + \mu^2\mathbf{F} + 2\mu\mathbf{C}. \tag{25}$$

Consequently:

1. ***Total variance inflation.*** *If*

$$|\mu| > \frac{2|\mathrm{tr}(\mathbf{C})|}{\mathrm{tr}(\mathbf{F})}, \tag{26}$$

then

$$\mathrm{tr}(\mathrm{Cov}(\hat{\mathbf{g}}_\mu)) > \mathrm{tr}(\mathrm{Cov}(\hat{\mathbf{g}}_0)). \tag{27}$$

2. ***Strict inflation in Lowner order.*** *If* $\mathbf{F} \succ 0$ *and*

$$|\mu| > \frac{2\|\mathbf{C}\|_2}{\lambda_{\min}(\mathbf{F})}, \tag{28}$$

then

$$\mathrm{Cov}(\hat{\mathbf{g}}_\mu) \succ \mathrm{Cov}(\hat{\mathbf{g}}_0). \tag{29}$$

3. ***Asymptotic quadratic dominance.*** *As* $|\mu| \to \infty$,

$$\mathrm{Cov}(\hat{\mathbf{g}}_\mu) - \mathrm{Cov}(\hat{\mathbf{g}}_0) = \mu^2\mathbf{F} + \mathcal{O}(|\mu|), \tag{30}$$

so the covariance penalty is asymptotically dominated by the quadratic term $\mu^2\mathbf{F}$.

Proof. Since $\mathbb{E}[\hat{\mathbf{g}}_\mu] = \mathbb{E}[\hat{\mathbf{g}}_0]$, the covariance difference equals the difference of the second-moment matrices:

$$\begin{aligned}\Delta\mathrm{Cov} :&= \mathrm{Cov}(\hat{\mathbf{g}}_\mu) - \mathrm{Cov}(\hat{\mathbf{g}}_0) \\ &= \mathbb{E}[\hat{\mathbf{g}}_\mu \hat{\mathbf{g}}_\mu^\top] - \mathbb{E}[\hat{\mathbf{g}}_0 \hat{\mathbf{g}}_0^\top]. \end{aligned} \tag{31}$$

Expanding $\hat{\mathbf{g}}_\mu = (A^* + \mu)\mathbf{g}$, we obtain

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{g}}_\mu \hat{\mathbf{g}}_\mu^\top] =& \mathbb{E}\Big[(A^*)^2 \mathbf{g}\mathbf{g}^\top\Big] + 2\mu\, \mathbb{E}\Big[A^* \mathbf{g}\mathbf{g}^\top\Big] + \\ & \mu^2\, \mathbb{E}\Big[\mathbf{g}\mathbf{g}^\top\Big]. \end{aligned} \tag{32}$$

Recognizing the first term as $\mathbb{E}[\hat{\mathbf{g}}_0 \hat{\mathbf{g}}_0^\top]$, the second as $2\mu\mathbf{C}$, and the third as $\mu^2\mathbf{F}$, we get

$$\Delta\mathrm{Cov} = \mu^2\mathbf{F} + 2\mu\mathbf{C}. \tag{33}$$

For the trace statement,

$$\mathrm{tr}(\Delta\mathrm{Cov}) = \mu^2 \mathrm{tr}(\mathbf{F}) + 2\mu\, \mathrm{tr}(\mathbf{C}). \tag{34}$$

Using $|2\mu\, \mathrm{tr}(\mathbf{C})| \le 2|\mu|\, |\mathrm{tr}(\mathbf{C})|$, we have

$$\mathrm{tr}(\Delta\mathrm{Cov}) \ge |\mu| \Big(|\mu| \mathrm{tr}(\mathbf{F}) - 2|\mathrm{tr}(\mathbf{C})| \Big). \tag{35}$$

Hence $\mathrm{tr}(\Delta\mathrm{Cov}) > 0$ whenever

$$|\mu| > \frac{2|\mathrm{tr}(\mathbf{C})|}{\mathrm{tr}(\mathbf{F})}. \tag{36}$$

For the Lowner-order statement, let $\mathbf{v}$ be any unit vector. Then

$$\mathbf{v}^\top \Delta\mathrm{Cov}\, \mathbf{v} = \mu^2 \mathbf{v}^\top \mathbf{F}\mathbf{v} + 2\mu\, \mathbf{v}^\top \mathbf{C}\mathbf{v}. \tag{37}$$

By Rayleigh-quotient bounds,

$$\mathbf{v}^\top \mathbf{F}\mathbf{v} \ge \lambda_{\min}(\mathbf{F}), \qquad |\mathbf{v}^\top \mathbf{C}\mathbf{v}| \le \|\mathbf{C}\|_2. \tag{38}$$

Therefore

$$\begin{aligned}\mathbf{v}^\top \Delta\mathrm{Cov}\, \mathbf{v} &\ge \mu^2 \lambda_{\min}(\mathbf{F}) - 2|\mu| \|\mathbf{C}\|_2 \\ &= |\mu| \Big(|\mu| \lambda_{\min}(\mathbf{F}) - 2\|\mathbf{C}\|_2 \Big). \end{aligned} \tag{39}$$

If

$$|\mu| > \frac{2\|\mathbf{C}\|_2}{\lambda_{\min}(\mathbf{F})}, \tag{40}$$

then $\mathbf{v}^\top \Delta\mathrm{Cov}\, \mathbf{v} > 0$ for every unit vector $\mathbf{v}$, which implies

$$\Delta\mathrm{Cov} \succ 0. \tag{41}$$

Finally, since

$$\Delta\mathrm{Cov} = \mu^2\mathbf{F} + 2\mu\mathbf{C}, \tag{42}$$

the linear term is lower order than the quadratic term as $|\mu| \to \infty$, proving

$$\Delta\mathrm{Cov} = \mu^2\mathbf{F} + \mathcal{O}(|\mu|). \tag{43}$$

□

Conclusion. In very high-dimensional models such as LVLMs, the Fisher term $\mathbf{F}$ can be large in aggregate, so the quadratic penalty $\mu^2\mathbf{F}$ may become significant even for modest $|\mu|$. While the exact threshold depends on how both $\mathbf{F}$ and $\mathbf{C}$ scale with dimension, the proposition makes clear that non-zero mean shifts become increasingly undesirable in large models because they inject variance in every parameter direction captured by the FIM. This degrades gradient signal-to-noise ratio and can materially destabilize optimization. This theoretical formulation strictly proves the necessity of maintaining a mathematically exact zero-mean advantage. It serves as the foundational justification for the sum-preserving normalization mechanism proposed in our PGPO framework (Equation 7), ensuring that the modulated advantages fundamentally conserve the $\mu = 0$ property while redistributing credit effectively.

H Monotonicity Analysis of $\tilde{\omega}_t$

Proposition G.1 (Monotonicity of Credit Assignment). *Given a generated sequence, let $\mathcal{S}_t$ be the visual dependency score of the t-th token. Assuming the scores of all other tokens in the sequence remain constant, the final modulated weight $\tilde{\omega}_t$ is monotonically non-decreasing with respect to $\mathcal{S}_t$.*

Proof. We prove this proposition by systematically analyzing the monotonicity of the four sequential transformations applied to $\mathcal{S}_t$: logarithmic compression, min-max normalization, threshold gating, and sum-preserving normalization.

Step 1: Logarithmic Compression. The first transformation applies a logarithmic dampening: $\tilde{\mathcal{S}}_t = \log(1 + \mathcal{S}_t)$. Assuming $\mathcal{S}_t \ge 0$, the partial derivative is:

$$\frac{\partial \tilde{\mathcal{S}}_t}{\partial \mathcal{S}_t} = \frac{1}{1 + \mathcal{S}_t} > 0 \tag{44}$$

Thus, $\tilde{\mathcal{S}}_t$ is strictly monotonically increasing with respect to $\mathcal{S}_t$.

Step 2: Min-Max Normalization. The second transformation standardizes the score:

$$I_t = \frac{\tilde{\mathcal{S}}_t - \min_j \tilde{\mathcal{S}}_j}{\max_j \tilde{\mathcal{S}}_j - \min_j \tilde{\mathcal{S}}_j + \epsilon} \tag{45}$$

Because the $\min$ and $\max$ operators evaluate the entire sequence, altering $\tilde{\mathcal{S}}_t$ may shift the sequence extrema. Let $m = \min_{j \ne t} \tilde{\mathcal{S}}_j$ and $M =$

$\max_{j \neq t} \tilde{\mathcal{S}}_j$, which act as fixed constants with respect to $\tilde{\mathcal{S}}_t$. We analyze I_t across three disjoint intervals:

1. **When** $\tilde{\mathcal{S}}_t \leq m$ (Token t is the minimum): $\min_j \tilde{\mathcal{S}}_j = \tilde{\mathcal{S}}_t$. The numerator becomes $\tilde{\mathcal{S}}_t - \tilde{\mathcal{S}}_t = 0$, yielding $I_t = 0$. The derivative is 0.

2. **When** $m < \tilde{\mathcal{S}}_t < M$ (Token t is an intermediate value): Here, $\min_j \tilde{\mathcal{S}}_j = m$ and $\max_j \tilde{\mathcal{S}}_j = M$. The derivative is:

$$\frac{\partial I_t}{\partial \tilde{\mathcal{S}}_t} = \frac{1}{M - m + \epsilon} > 0 \tag{46}$$

3. **When** $\tilde{\mathcal{S}}_t \geq M$ (Token t is the maximum): Here, $\max_j \tilde{\mathcal{S}}_j = \tilde{\mathcal{S}}_t$ and $\min_j \tilde{\mathcal{S}}_j = m$. Applying the quotient rule:

$$\begin{aligned} \frac{\partial I_t}{\partial \tilde{\mathcal{S}}_t} &= \frac{\partial}{\partial \tilde{\mathcal{S}}_t} \left(\frac{\tilde{\mathcal{S}}_t - m}{\tilde{\mathcal{S}}_t - m + \epsilon} \right) \\ &= \frac{\epsilon}{(\tilde{\mathcal{S}}_t - m + \epsilon)^2} > 0 \end{aligned} \tag{47}$$

Across all intervals, $\frac{\partial I_t}{\partial \tilde{\mathcal{S}}_t} \geq 0$. Hence, I_t is monotonically non-decreasing with respect to $\tilde{\mathcal{S}}_t$.

Step 3: Threshold-Gating Function. The piecewise gating function computes the base weight $\omega_t = \omega(I_t)$. We analyze its piecewise derivatives and boundary continuity:

1. **When** $I_t < \tau$: $\omega_t = \frac{I_t}{\tau + \epsilon} \implies \frac{\partial \omega_t}{\partial I_t} = \frac{1}{\tau + \epsilon} > 0$.

2. **When** $I_t \geq \tau$: $\omega_t = 1 + \beta \frac{I_t - \tau}{1 - \tau + \epsilon} \implies \frac{\partial \omega_t}{\partial I_t} = \frac{\beta}{1 - \tau + \epsilon} \geq 0$ (strictly positive for $\beta > 0$).

3. **Boundary at** $I_t = \tau$: The right-hand evaluation is $\omega(\tau) = 1$. The left-hand limit is $\lim_{I_t \to \tau^-} \omega(I_t) = \frac{\tau}{\tau + \epsilon}$. Since $\epsilon > 0$, the left limit is strictly less than 1. This positive jump discontinuity guarantees that the function strictly preserves the non-decreasing property as it crosses the threshold.

Consequently, ω_t is monotonically non-decreasing with respect to I_t.

Step 4: Sum-Preserving Renormalization. The final step enforces mass conservation:

$$\tilde{\omega}_t = \omega_t \cdot \frac{T}{\sum_{j=1}^{T} \omega_j} = \frac{T \cdot \omega_t}{\omega_t + S} \tag{48}$$

where T is the sequence length and $S = \sum_{j \neq t} \omega_j > 0$ is a fixed positive constant. Taking the derivative with respect to ω_t:

$$\frac{\partial \tilde{\omega}_t}{\partial \omega_t} = \frac{T \cdot S}{(\omega_t + S)^2} > 0 \tag{49}$$

Thus, $\tilde{\omega}_t$ is strictly monotonically increasing with respect to ω_t. □

Conclusion. The overall mapping from $\mathcal{S}_t$ to $\tilde{\omega}_t$ is a composition of functions that are differentiable and non-decreasing almost everywhere, with any discontinuities being strictly positive jumps. By the properties of monotonic compositions, the chain strictly preserves order. Therefore, $\tilde{\omega}_t$ is monotonically non-decreasing with respect to $\mathcal{S}_t$. In the context of credit assignment, this mathematical property guarantees that if a specific token demonstrates stronger visual grounding (i.e., higher $\mathcal{S}_t$), the PGPO framework will reliably assign it an equal or greater advantage modulation weight $\tilde{\omega}_t$, ensuring stable and interpretable optimization.

I Rank-Preserving Property of $\tilde{\omega}_t$

Proposition H.1 (Intra-Sequence Rank Preservation). *Given a generated sequence, the PGPO modulation mechanism strictly preserves the relative ordinal importance of the tokens. Specifically, for any two tokens A and B within the same sequence, if their raw visual dependency scores satisfy $\mathcal{S}_A > \mathcal{S}_B$, then their final modulated weights satisfy $\tilde{\omega}_A > \tilde{\omega}_B$.*

Proof. Let A and B be two distinct tokens in a fixed sequence such that $\mathcal{S}_A > \mathcal{S}_B$. Because both tokens belong to the same sequence, the sequence-level statistics—specifically the unnormalized minimum m, maximum M, and the pre-normalized sum $S_{\text{total}} = \sum_{j=1}^{T} \omega_j$—are fixed, shared constants. We evaluate the strict preservation of the inequality through the four transformations:

1. Logarithmic Compression: Since $f(x) = \log(1 + x)$ is strictly monotonically increasing for $x \geq 0$, we have:

$$\mathcal{S}_A > \mathcal{S}_B \implies \tilde{\mathcal{S}}_A > \tilde{\mathcal{S}}_B \tag{50}$$

2. Min-Max Normalization: The normalization applies a linear transformation:

$$I_A = \frac{\tilde{\mathcal{S}}_A - m}{M - m + \epsilon}, \quad I_B = \frac{\tilde{\mathcal{S}}_B - m}{M - m + \epsilon} \tag{51}$$

Because the denominator $(M - m + \epsilon) > 0$ is identical for both tokens, the order is strictly preserved: $I_A > I_B$.

3. Threshold-Gating Function: For $\beta > 0$, the piecewise function $\omega(I)$ consists of two linear segments with strictly positive slopes, connected by a positive upward jump at $I = \tau$. Because the function is globally strictly monotonically increasing across its domain $[0, 1]$, we obtain:

$$I_A > I_B \implies \omega_A > \omega_B \tag{52}$$

4. Sum-Preserving Renormalization: The final step applies sequence-level scaling:

$$\tilde{\omega}_A = \omega_A \cdot \frac{T}{S_{\text{total}}}, \quad \tilde{\omega}_B = \omega_B \cdot \frac{T}{S_{\text{total}}} \tag{53}$$

Since the scalar multiplier $\frac{T}{S_{\text{total}}} > 0$ is a shared constant, the inequality is maintained: $\tilde{\omega}_A > \tilde{\omega}_B$. □

Conclusion. Through all sequential operations, the relationship $\mathcal{S}_A > \mathcal{S}_B \implies \tilde{\omega}_A > \tilde{\omega}_B$ holds strictly true. Therefore, the proposed modulation mechanism mathematically guarantees consistency between a token's empirical visual grounding and its allocated advantage weighting.

J Analysis of the Training Dataset

We use `ViRL39K` (Wang et al., 2025a) as the sole training source for reinforcement learning, which is under MIT licenses. Its construction matches our objective of learning dependable multimodal reasoning under verifiable supervision.

Breadth of problem domains. The dataset contains roughly 39K instances spanning mathematics, physics, chemistry, biology, and chart-centric interpretation. This cross-domain coverage exposes the policy to heterogeneous reasoning patterns and reduces over-specialization to narrow distributions.

Deterministic rewardability. Each synthesized sample is paired with a unique, machine-checkable reference answer. This property is central to Reinforcement Learning with Verifiable Rewards (RLVR), where stable optimization depends on objective and automated reward assignment rather than subjective model-as-judge scoring.

K Analysis of Evaluation Benchmarks

Our evaluation protocol uses seven benchmarks to measure complementary abilities: domain-intensive mathematical reasoning, general logic, and cross-discipline multimodal understanding. These datasets are licensed under either the Apache or MIT open-source licenses.

- **DynaMath** (Zou et al., 2024): tests robustness by programmatically perturbing numeric values and functional plots in seed questions, helping distinguish memorization from true generalization. We evaluate on *sample variant1*.
- **Geo3k** (Lu et al., 2021): a large high-school geometry benchmark with formal annotations, suitable for evaluating symbolic and interpretable solution behavior.
- **MathVerse** (Zhang et al., 2024): provides six controlled variants that shift information between text and diagram, enabling fine-grained analysis of visual reliance. We use the testmini multi-choice subset.
- **MATH-Vision** (Wang et al., 2024a): sourced from authentic contests (e.g., AMC, Math Kangaroo), spanning 16 subjects and five difficulty levels. We report results on its full test split.
- **MMK12** (Meng et al., 2025): focuses on K–12 multimodal math and is used to assess foundational reasoning competence.
- **LogicVista** (Xiao et al., 2024): evaluates inductive, deductive, numerical, spatial, and mechanical reasoning across diverse visual formats, offering a broad test of general logical cognition.
- **MMMU-Pro** (Yue et al., 2024): an upgraded MMMU variant designed to suppress text-only shortcuts via stronger option design and vision-centric settings, thereby better testing joint visual-textual reasoning in academic contexts.

L LLM Usage Statement

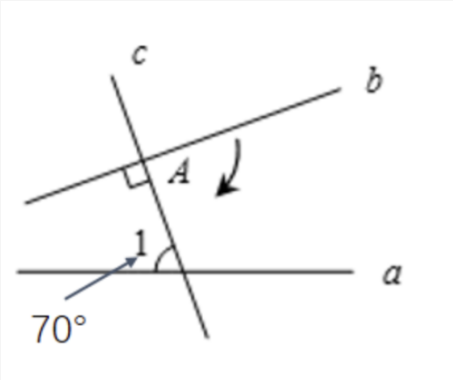
In preparing this manuscript, we used a large language model (LLM) solely to refine language expression, such as correcting grammar and improving sentence fluency. It was not involved in developing the scientific ideas, conducting the analyses,

or producing the core technical content. The authors retain full responsibility for all intellectual contributions and for the final version of the paper.

M Case Study

To illustrate how PGPO improves reasoning, we present three representative case studies. We contrast typical failure patterns of the baseline with the correct reasoning paths produced by PGPO-7B. In all three cases, PGPO-7B answers correctly in all eight sampled generations, demonstrating strong stability and consistency. These examples highlight the benefit of token-level credit assignment. While a uniform training signal can cause key errors in visual understanding or logical inference, PGPO emphasizes visually grounded critical tokens and guides reasoning toward the correct answer.

Case 1



Question: As shown in the figure, If the line b is parallel to the line a, the line b can be rotated () clockwise around the point A

`Source: MathVerse, id: 74`

Choices: (A) 70° (B) 50° (C) 30° (D) 20°

Ground truth: (D) 20°

DAPO:

<think>

Given that line b is parallel to line a, and an angle measure between them is a given exterior angle at A. Specifically, the given angle adjacent to ∠ 1 between line b and the horizontal line is 70°. Since b ∥ a, the angle ∠ 1 can be calculated using the supplementary relationship with the adjacent given angle. So we need to find the rotation angle in order to make b ∥ a. Here the supplementary relation: ∠ 1 + supplementary to it = 180°; So rotation = 180° - 70°;

</think>

The correct answer is: $\boxed{\text{B}}$ Wrong

PGPO (Ours):

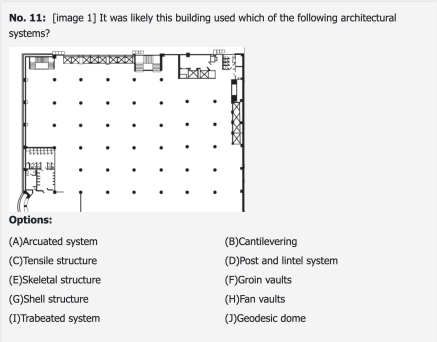
<think>

We are required to determine how much line b can rotate clockwise around point A for the lines to maintain the parallel condition. Identify the angles: The angle at point A where line b intersects with a horizontal line is given 70°. Understand the rotation requirement: We know currently angle 1 is 70°. For b to parallel to a, rotation is needed. The current angle effectively dictates that the rotation needed is to move from 70° to a position keeping b parallel to a. Identify angle at rotation: Given current alignment requires b shifted from 70° configuration: Notice the perpendicular creating 90°, angle rotation at A: Since it currently reads 70°, and b needs to realign, rotation = 90° - 70° = 20°. Thus, b needs to rotate 20° clockwise around A.

</think>

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

Case 2



Question: (In the image)

`Source: MMMU-Pro, id: 19`

Choices: (In the image)

Ground truth: (E) Skeletal structure

DAPO:

<think>

The image shows a floor plan with a grid system, where there are supporting columns or beams at regular intervals. This kind of arrangement suggests a structure where vertical posts (columns) are connected horizontally by lintels (beams). Given this layout, the architectural system that fits is the Post and lintel system, which is typically used when there's a need to support a structure with evenly distributed vertical posts and horizontal lintels. So, among the given options: (A) Arcuated system - deals with curving structures. (B) Cantilevering - involves unsupported overhang. (C) Tensile structure - relies on cables for support. (D) Post and lintel system - aligns perfectly with evenly distributed supports shown. (E) Skeletal structure - involves a framework. (F) Groin vaults - require pointed arches connecting horizontal and vertical systems. (G) Shell structure - involves curved exterior forms. (H) Fan vaults - are elaborately used in medieval architecture. (I) Trabeated system - resembles the post-and-lintel system.

</think>

The correct answer is: $\boxed{\text{D}}$ Wrong

PGPO (Ours):

<think>

The image shows a floor plan with a regular grid of columns and beams. This suggests a structure where supports are aligned in a grid, which is typical for a skeletal structure where the main load-bearing elements are vertical columns and horizontal beams in a grid pattern to support the floor and roof. There is no indication of arches, vaults, or any specific architectural system mentioned in options (A, F, H, I, J). The regular column spacing and arrangement are indicative of a skeletal structure where elements are connected at joints to form a framework that supports the building's structure. Therefore, the most fitting option here is (E) Skeletal structure. </think>

The correct answer is: $\boxed{\text{E}}$ Correct

Case 3

Question: (In the image)

`Source: MMMU-Pro, id: 22`

Choices: (In the image)

Ground truth: (H) A book cover

DAPO:

<think>

The image shows a beautifully detailed piece that appears to be a flat, decorative work adorned with gemstones and a central figure, suggesting it is meant for display. Given the ornate and decorative style, along with the inclusion of a central iconographic element (the figure of Jesus on the cross), this item likely served a ceremonial or devotional purpose in a Christian context. Reviewing the options: (A) An icon: Icons are religious artworks primarily used as objects of veneration, matching the devotional and symbolic nature of the artwork. (B) An altarpiece: Typically a three-dimensional piece, this seems more like a flat display. (C) A sculpture: Not applicable because the image isn't three-dimensional. (D), (F), (J): These are also three-dimensional forms, unsuitable for the image. (E), (G), (H), (I): These forms are not as typical of elaborate decorative flat devotional pieces adorned with jewelry. Thus, the artwork's nature as a decorative yet religious symbol points toward (A) an icon.

</think>

The correct answer is: $\boxed{\text{A}}$ Wrong

PGPO (Ours):

<think>

The image shows a decorative, gem-encrusted piece with a central religious icon of a crucifix, surrounded by ornamental details. This type of artifact is typically used as a cover for a book, especially a holy book or a Bible, given its stylistic and symbolic representation. An icon typically refers to a representation of a sacred subject, often used in Christian religious contexts, which matches the imagery on the piece. The luxurious nature with gems and the central iconography suggest it serves as a protective and valuable cover, often for religious texts like the Bible. Given the options: (A) an icon: This usually refers to a small religious image or painting, not necessarily the work shown here. (B) an altarpiece: Altarpieces are larger and mounted on an altar, which doesn't match the size and style of the object. (C) a sculpture: The piece is not a three-dimensional sculpture. (D) a vase design: The piece is not designed to be a vase. (E) a door ornament: This piece is not a door ornament. (F) a fresco: Frescoes are painted on walls, which this is not. (G) a tapestry: While it has ornate detail, it seems too precious to be a tapestry. (H) a book cover: The ornamented and iconographic style fits a book cover. (I) a purse lid: This piece is decorative and religious, not a purse lid. (J) a mural: Murals are typically on larger surfaces like walls, not these kinds of objects. The most fitting description here is a book cover because of its rich iconographic nature and the way it would be used to cover a book. </think>

The correct answer is: $\boxed{\text{H}}$ Correct