# DSE-GRPO: DIFFERENTIAL SEMANTIC ENTROPY FOR DEBIASED NOISY-ROLLOUT POLICY OPTIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reinforcement learning (RL) post-training for vision–language models (VLMs) increasingly relies on noisy rollouts to encourage exploration and improve robustness. However, mixing rollouts generated from clean and perturbed images introduces a hidden off-policy mismatch: noise can change the semantic interpretation of the input, yielding biased and high-variance policy gradients when these trajectories are optimized together under GRPO. We propose DSE-GRPO, an uncertainty-aware variant of GRPO that explicitly measures and suppresses noise-induced semantic drift. Our key idea is Differential Semantic Entropy (DSE): we estimate semantic entropy over response clusters for clean and noisy rollout groups separately, and use their non-negative gap as a differential uncertainty signal. DSE-GRPO then adaptively down-weights the advantages of noisy rollouts with large DSE, filtering harmful off-policy updates while retaining beneficial exploratory trajectories. We provide theoretical analysis showing that DSE-based reweighting attenuates off-policy bias and stabilizes optimization dynamics, while slowing policy-entropy collapse to preserve exploration. Empirically, RL-tuning with Geometry3K and MMK12 shows that DSE-GRPO consistently outperforms vanilla GRPO and NoisyRollout baselines, improving the average accuracy of Qwen2.5-VL-7B and Qwen3-VL-8B by 8.0 and 1.9 percentage points, respectively, across five out-of-domain benchmarks, with markedly smoother training dynamics.

## 1 INTRODUCTION

Reinforcement learning (RL) post-training supports reasoning in vision–language models (VLMs) through verifiable rewards. Group Relative Policy Optimization (GRPO) (Guo et al., 2025; Shao et al., 2024) estimates advantages from grouped rollouts without the value model used in PPO (Schulman et al., 2017). Robustness to visual corruption, distribution shifts, and spurious perceptual cues is also important when these models are deployed.

NoisyRollout (Liu et al., 2025) promotes robustness and exploration by jointly optimizing responses generated from clean and stochastically distorted images. These noisy trajectories are *off-policy with respect to the clean-input objective*: perturbations can change the semantic interpretation of the input and bias updates toward the perturbed-input objective. The effect on aggregate gradient variance depends on how the updates covary.

Selective reweighting must address two concerns. *Advantage-signal pollution* can arise when a perturbation-induced reward change affects an entire response, including tokens unaffected by the perturbation. Recent work on entropy control (Petrenko et al., 2026) and vision-aware credit assignment (Jin et al., 2026) further motivates separating exploration from reliable visual grounding. Neither concern alone establishes a causal link between noise, gradient variance, and entropy collapse. Discarding every uncertain group could also remove useful exploratory responses. SEED-GRPO (Chen et al., 2025b) uses absolute semantic entropy to attenuate uncertain groups, but this signal cannot distinguish existing uncertainty from an increase under noise.

We propose **Differential Semantic Entropy GRPO (DSE-GRPO)** to regulate noisy updates using the change in answer uncertainty. For each prompt, we compute semantic entropy (Farquhar et al., 2024) over clean and noisy rollouts separately and define DSE as their non-negative difference. DSE supplies an additional attenuation factor for noisy advantages when noise increases answer dispersion. The clean group provides a reference for added uncertainty, although the entropy gap



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

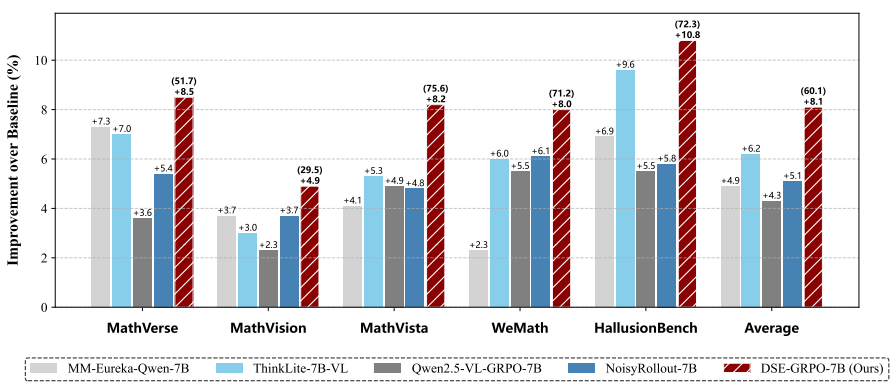



Figure 1: Accuracy improvement over Qwen2.5-VL-7B-Instruct on 5 out-of-domain benchmarks, covering both visual reasoning tasks (from MathVerse to WeMath) and a visual perception task (HallusionBench). Qwen2.5-VL-GRPO-7B, NoisyRollout-7B and DSE-GRPO-7B are fine-tuned by ourselves (denoted with ‡) with only 2.1K training samples from Geometry3K. The exact accuracy of DSE-GRPO-7B is annotated above each corresponding bar in parentheses.

does not capture every form of off-policy contamination. Coherent answer shifts and reasoning errors can occur without increasing entropy (Appendix B).

We evaluate DSE-GRPO on RL fine-tuning with Geometry3K (Lu et al., 2021) and MMK12 (Meng et al., 2025), and assess generalization on five out-of-domain benchmarks spanning visual reasoning (MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), MathVista (Lu et al., 2023), WeMath (Qiao et al., 2025)) and visual perception (HallusionBench (Guan et al., 2024)), with results in Tables 1 and 2. DSE-GRPO improves the average over the Qwen2.5-VL-7B and Qwen3-VL-8B base models by 8.0 and 1.9 percentage points, respectively. Qwen2.5-VL-3B comparisons test the method at a smaller model scale. Training curves for Qwen2.5-VL show smoother accuracy dynamics and slower policy-entropy decay in the reported settings.

Our contributions are:

- We identify a structural limitation of noisy-rollout GRPO training: mixing clean and noisy trajectories can change the effective optimization target when perturbations alter answer semantics.
- We propose **DSE-GRPO**, which introduces **differential semantic entropy** to quantify noise-induced uncertainty increments and uses it to adaptively down-weight noisy rollouts during policy optimization.
- We report average accuracy gains across training datasets and model scales, and use controlled ablations and training curves to assess the differential term, rollout allocation, and exploration behavior.

## 2 Related Work

**Vision–language models.** Image–text alignment in CLIP (Radford et al., 2021) and multimodal instruction tuning in InstructBLIP (Dai et al., 2023) underpin later VLMs such as InternVL (Chen et al., 2024), Qwen2.5-VL (Bai et al., 2025b), and Qwen3-VL (Bai et al., 2025a). Recent work also learns how to acquire visual evidence: Foveated Reasoner (Min et al., 2026) combines selective high-resolution viewing with RL. DSE-GRPO instead regulates updates from already sampled clean and perturbed observations.

**RL for visual reasoning.** GRPO (Shao et al., 2024; Guo et al., 2025) enables reasoning-oriented fine-tuning without a separate value model. GSPO (Zheng et al., 2025a), DAPO (Yu et al., 2025), and SRPO (Wan et al., 2025) study alternative optimization and rollout strategies. In 2026, REPO and ADAPO (Petrenko et al., 2026) explicitly regulate policy entropy through advantage modification and adaptive clipping. VEPO (Jin et al., 2026) combines token entropy with visual sensitivity for

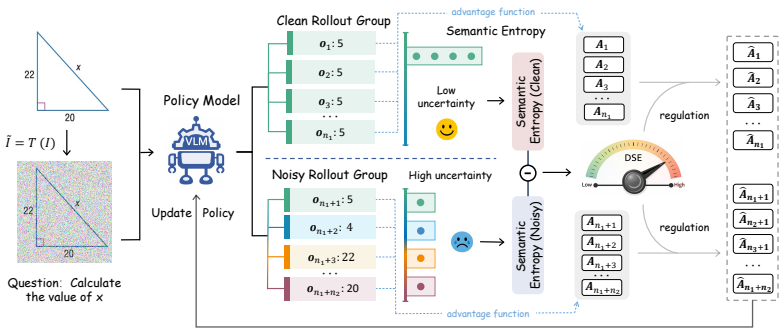



Figure 2: Overview of DSE-GRPO. Responses to clean image $I$ and perturbed image $\tilde{I}$ are clustered separately to compute semantic entropy. The positive entropy gap provides an additional weight for noisy advantages. The figure indexes clean rollouts by $j = 1, \ldots, n_1$ and noisy rollouts by $k = n_1 + 1, \ldots, n_1 + n_2$.

credit assignment. DSE-GRPO operates at the rollout-group level using final-answer uncertainty, rather than token selection or a direct policy-entropy target.

**Noisy rollouts.** NoisyRollout (Liu et al., 2025) mixes clean and distorted-image responses. NC-GRPO (Jerge et al., 2026) instead perturbs prompt hidden states to diversify rollouts while retaining the original RL objective. DSE-GRPO retains image perturbations and changes advantage weights according to the clean–noisy entropy gap. The location of perturbation and the weighting of its resulting updates are distinct design choices.

**Semantic-entropy weighting.** SEED-GRPO (Chen et al., 2025b) uses semantic entropy (Farquhar et al., 2024) to attenuate uncertain groups. DSE-GRPO adds a differential factor for noisy responses, separating increased answer dispersion from uncertainty already present in the clean group. This signal does not detect every semantic shift. Section 4.1 compares the methods; Section 4.2 isolates the differential factor.

## 3 Differential Semantic Entropy GRPO

### 3.1 Preliminaries

**GRPO.** For an image–question pair $(I, q)$, GRPO (Guo et al., 2025; Shao et al., 2024) samples $G$ responses from the rollout policy $\pi_{\theta_{\text{old}}}$. Response $o_i$ receives reward $R_i \in \{0, 1\}$ according to whether its extracted final answer matches the target. The group statistics and relative advantage are

$$\mu_R = \frac{1}{G}\sum_{i=1}^{G} R_i, \qquad \sigma_R^2 = \frac{1}{G}\sum_{i=1}^{G}(R_i - \mu_R)^2,$$

$$A_i = \begin{cases} (R_i - \mu_R)/\sigma_R, & \sigma_R > 0, \\ 0, & \sigma_R = 0. \end{cases}$$

The zero-variance group supplies no relative reward signal. PPO-style clipping (Schulman et al., 2017) limits the probability-ratio update; following NoisyRollout, we omit a reference-policy KL penalty.

**Semantic entropy.** We estimate semantic entropy (Farquhar et al., 2024) from final-answer frequencies. For a nonempty response group $\mathcal{O}$, remove exact-duplicate response strings to obtain $m$ retained responses and partition them into $K$ clusters by normalized final answer. Distinct reasoning strings with the same answer remain separate observations in the same cluster. Define

$$p_k = \frac{|C_k|}{m}, \qquad \mathrm{SE}(\mathcal{O}) = -\sum_{k=1}^{K} p_k \log p_k, \qquad \sum_{k=1}^{K} p_k = 1. \tag{1}$$

We use natural logarithms and $0 \log 0 = 0$. Thus $0 \le \mathrm{SE}(\mathcal{O}) \le \log K \le \log m$; entropy is zero when all retained answers agree. Unlike sequence-likelihood weighting, this estimator counts answers

directly. SEED-GRPO (Chen et al., 2025b) attenuates advantages using the group's absolute semantic entropy.

### 3.2 DSE-GRPO Algorithm

For each $(I, q)$, draw a perturbed image $\tilde{I}$ and sample clean and noisy responses separately:

$$\mathcal{O}_c = \{o_j\}_{j=1}^{n_1}, \qquad o_j \sim \pi_{\theta_{\text{old}}}(\cdot \mid I, q),$$
$$\mathcal{O}_n = \{o_k\}_{k=n_1+1}^{n_1+n_2}, \quad o_k \sim \pi_{\theta_{\text{old}}}(\cdot \mid \tilde{I}, q).$$

Rewards are pooled over $G = n_1 + n_2$ responses to compute $A_i$. Entropy is computed separately within each group (Figure 2). For $n_1, n_2 > 1$, define

$$\text{SE}_c = \frac{\text{SE}(\mathcal{O}_c)}{\log n_1}, \qquad \text{SE}_n = \frac{\text{SE}(\mathcal{O}_n)}{\log n_2}, \qquad \text{SE}_c, \text{SE}_n \in [0, 1]. \tag{2}$$

The denominators use the original rollout counts; deduplication affects only the entropy estimate. A singleton group is assigned normalized entropy zero. The zero-noisy-rollout ablation uses the clean-only GRPO baseline. The differential signal is

$$\text{DSE} = \max\{0, \text{SE}_n - \text{SE}_c\} \in [0, 1]. \tag{3}$$

We use the linear attenuation maps $f(x) = g(x) = 1 - x$ with $\alpha, \beta \in [0, 1]$, giving

$$\begin{aligned} \hat{A}_j &= A_j(1 - \alpha \text{SE}_c), && j = 1, \ldots, n_1, \\ \hat{A}_k &= A_k(1 - \alpha \text{SE}_n)(1 - \beta \text{DSE}), && k = n_1 + 1, \ldots, n_1 + n_2. \end{aligned} \tag{4}$$

Both groups retain absolute-entropy weighting; only noisy responses receive the differential factor. At fixed noisy entropy, this factor decreases linearly with DSE. It does not certify trajectory correctness. Advantages and entropy weights are held fixed during policy optimization.

To match the `token-mean` aggregation, let $\mathcal{B}$ contain the clean and noisy responses in an optimization batch. Response $i$ uses its generating observation $I_i \in \{I, \tilde{I}\}$ and question $q_i$. Let $M_{it}$ indicate a valid response token. Define

$$N_{\mathcal{B}} = \sum_{i,t} M_{it} > 0, \qquad h_{it} = (I_i, q_i, o_{i,<t}), \qquad \rho_{it}(\theta) = \frac{\pi_\theta(o_{it} \mid h_{it})}{\pi_{\theta_{\text{old}}}(o_{it} \mid h_{it})}.$$

The clipped objective is

$$\mathcal{J}(\theta) = \mathbb{E}_{\mathcal{B}}\left[\frac{1}{N_{\mathcal{B}}} \sum_{i,t} M_{it}\, \ell_\epsilon\Big(\rho_{it}(\theta), \hat{A}_i\Big)\right], \tag{5}$$
$$\ell_\epsilon(\rho, A) = \min\{\rho A, \text{clip}(\rho, 1 - \epsilon, 1 + \epsilon)A\}.$$

The expectation includes prompts, image perturbations, and responses sampled from the rollout policy. Both groups share the valid-token denominator. Each ratio conditions on the observation that generated its token; it does not correct the noisy-input distribution to the clean-input distribution. Algorithm 1 summarizes the training procedure.

### 3.3 Analysis of Uncertainty Reweighting

At the rollout policy, let $u_i$ be response $i$'s unweighted score-function contribution, including the objective's aggregation factor. With the multiplier $w_i$ from Eq. (4), define

$$G_0 = \sum_i u_i, \qquad G_D = \sum_i w_i u_i.$$

Thus $G_0$ removes only the entropy weights from the same surrogate. We fix the policy parameters, assume finite second moments, and take expectations over sampled training batches. Advantages and weights are detached during differentiation, but need not be statistically independent. Appendix A.1

**Algorithm 1** Training procedure of DSE-GRPO

**Input:** policy $\pi_\theta$, dataset $\mathcal{D}$, rollout counts $n_1, n_2 > 0$, coefficients $\alpha, \beta \in [0, 1]$, clip width $\epsilon > 0$, perturbation schedule $T_{\alpha_t}$.
**for** each rollout iteration $t$ **do**
  Freeze $\theta_{\text{old}} \leftarrow \theta$ and sample a batch $(I, q) \sim \mathcal{D}$
  Draw perturbed observations $\tilde{I} \leftarrow T_{\alpha_t}(I)$
  Sample $n_1$ clean responses from $\pi_{\theta_{\text{old}}}(\cdot \mid I, q)$ and $n_2$ noisy responses from $\pi_{\theta_{\text{old}}}(\cdot \mid \tilde{I}, q)$
  Compute rewards $R_i$ and pooled advantages $A_i$; set $A_i = 0$ if $\sigma_R = 0$
  Form answer-frequency clusters separately in $\mathcal{O}_c$ and $\mathcal{O}_n$
  Compute $\text{SE}_c, \text{SE}_n$ and $\text{DSE} = \max\{0, \text{SE}_n - \text{SE}_c\}$
  Set $\hat{A}_j \leftarrow A_j(1 - \alpha \text{SE}_c)$ for clean responses
  Set $\hat{A}_k \leftarrow A_k(1 - \alpha \text{SE}_n)(1 - \beta \text{DSE})$ for noisy responses
  Detach advantages and weights; compute the token probability ratios in Eq. (5)
  Update $\theta$ with AdamW by minimizing $\mathcal{L}(\theta) = -\mathcal{J}(\theta)$, keeping $\theta_{\text{old}}$ fixed
**end for**

gives the assumptions and proofs; these are local score-function results, not guarantees for multiple clipped AdamW updates.

**Analysis 1: Bias depends on alignment.**

For a fixed clean-input target gradient $G^\star$, define the original bias and its change:

$$\Delta = \mathbb{E}G_0 - G^\star, \qquad B = \mathbb{E}(G_D - G_0).$$

The exact change in squared bias is

$$\|\mathbb{E}G_D - G^\star\|^2 - \|\mathbb{E}G_0 - G^\star\|^2 = 2\langle \Delta, B \rangle + \|B\|^2. \tag{6}$$

Bias decreases precisely when the right-hand side is negative. Bounded, rollout-dependent weights do not ensure this alignment or an unbiased update.

**Analysis 2: Attenuation and update perturbation.**

For $\alpha, \beta \in [0, 1]$, the linear weights satisfy

$$\delta = \alpha + \beta - \alpha\beta, \qquad 1 - \delta \le w_i \le 1.$$

Consequently,

$$\|w_i u_i\| \le \|u_i\|, \qquad \|G_D - G_0\| \le \delta \sum_i \|u_i\|. \tag{7}$$

These bounds control individual contributions and the change from the unweighted update. They do not order aggregate variances: reweighting can remove cancellation between opposing contributions.

**Analysis 3: Local entropy-change bound.**

Let $H(\theta)$ be policy entropy averaged over a fixed context distribution. If $\nabla H$ is $L_H$-Lipschitz on every update segment, $\theta^+ = \theta + \eta G_D$ with $\eta \ge 0$ satisfies

$$\left|\mathbb{E}[H(\theta^+) - H(\theta)]\right| \le \eta \|\nabla H(\theta)\| \mathbb{E}\|G_D\| + \frac{L_H \eta^2}{2} \mathbb{E}\|G_D\|^2. \tag{8}$$

The bound retains the second-order remainder. It does not imply a sign for entropy change or preservation of exploration.

*Remark* 3.1 (Scope of differential selectivity). The weights decrease with the positive answer-entropy gap; the bounds above concern their attenuation, not detection of harmful trajectories. Coherent wrong answers and incorrect reasoning hidden by answer agreement can have zero DSE (Appendix B). Bias, variance, and exploration improvements therefore require evidence beyond bounded weights.

## 4 Experiments

**Training data and models.** Qwen2.5-VL-3B/7B-Instruct (Bai et al., 2025b) are trained separately on Geometry3K (Lu et al., 2021) (2.1K samples) and MMK12 (Meng et al., 2025) (6.4K samples).

Table 1: Full comparison on five OOD benchmarks (accuracy, %). Blue and red data sizes denote SFT and RL. Bold and underline mark the highest and second-highest reported scores in each column, including ties. DSE-GRPO rows are shaded. Geometry3K and MMK12 contain 2.1K and 6.4K training samples; Qwen3-VL-8B uses both. MVerse, MVis, MVista, WeM, and Hallu denote MathVerse, MathVision, MathVista, WeMath, and HallusionBench; Avg. is the reported five-benchmark mean.

| Model | Data Size | MVerse | MVis | MVista | WeM | Hallu | Avg. |
|---|---|---|---|---|---|---|---|
| *Open-source* | | | | | | | |
| InternVL2.5-8B-Instruct | - | 39.5 | 19.7 | 64.4 | - | 67.3 | - |
| InternVL3-8B | - | 39.8 | 29.3 | 71.6 | 37.1 | 49.9 | 45.5 |
| LLaVA-OneVision-7B | - | 26.2 | - | 63.2 | - | 48.4 | - |
| Kimi-VL-16B | - | 44.9 | 21.4 | 68.7 | - | 66.2 | - |
| URSA-8B | - | 45.7 | 26.2 | 59.8 | - | - | - |
| Mulberry-7B | - | - | - | 63.1 | - | - | - |
| *R1-related (reinforcement learning with verifiable reward)* | | | | | | | |
| R1-VL-7B | 260K+10K | 40.0 | 24.7 | 63.5 | - | - | - |
| R1-OneVision-7B | 155K+10K | 46.1 | 22.5 | 63.9 | 62.1 | 65.6 | 52.0 |
| OpenVLThinker-7B | 35K+15K | 48.0 | 25.0 | 71.5 | 67.8 | 70.8 | 56.6 |
| Vision-R1-7B | 200K+10K | 47.8 | - | 71.3 | - | - | - |
| MM-Eureka-Qwen-7B | 15K | 50.5 | 28.3 | 71.5 | 65.5 | 68.3 | 56.8 |
| ADORA-7B | 2.1K | 50.1 | 27.6 | 71.1 | 67.1 | 53.1 | 53.8 |
| ThinkLite-7B-VL | 11K | 50.2 | 27.6 | 72.7 | 69.2 | 71.0 | 58.1 |
| VLAA-Thinker-7B | 25K | 49.9 | 26.9 | 68.8 | 67.9 | 68.6 | 56.4 |
| NoisyRollout | 2.1K | 48.6 | 28.3 | 72.2 | 69.0 | 71.2 | 57.9 |
| NoisyRollout | 6.4K | 50.1 | 29.1 | 73.4 | 69.5 | 69.4 | 58.3 |
| SEED-GRPO | 2.1K | 48.8 | 28.7 | 72.6 | 69.3 | 71.1 | 58.1 |
| Qwen2.5-VL-7B-Instruct | - | 43.2 | 24.6 | 67.4 | 63.2 | 61.4 | 52.0 |
| + Vanilla GRPO | 2.1K | 46.8 | 26.9 | 72.3 | 68.7 | 70.5 | 57.0 |
| + DSE-GRPO | 2.1K | 50.6 | 29.3 | 74.0 | 70.7 | 72.1 | 59.4 |
| + Vanilla GRPO | 6.4K | 47.9 | 28.9 | 72.9 | 69.2 | 70.8 | 57.9 |
| + DSE-GRPO | 6.4K | 51.7 | 29.5 | 75.6 | 71.2 | 71.2 | 60.0 |
| Qwen3-VL-8B-Instruct | - | <u>60.8</u> | <u>52.9</u> | <u>77.9</u> | <u>79.4</u> | <u>75.0</u> | <u>69.2</u> |
| + DSE-GRPO | 2.1K+6.4K | **63.4** | **54.6** | **79.5** | **82.5** | **75.4** | **71.1** |

Qwen3-VL-8B-Instruct (Bai et al., 2025a) is trained on their combination, denoted 2.1K+6.4K. Following NoisyRollout (Liu et al., 2025), multiple-choice training items are reformulated as open-ended queries to reduce answer guessing.

The comparisons in Tables 1 and 2 include InternVL2.5 (4B/8B) (Chen et al., 2025c), InternVL3 (2B/8B) (Zhu et al., 2025), LLaVA-OneVision-7B (Li et al., 2024), Kimi-VL-16B (Team et al., 2025), URSA-8B (Luo et al., 2025), Mulberry (2B/7B) (Yao et al., 2024), R1-VL (2B/7B) (Zhang et al., 2025), R1-OneVision (3B/7B) (Yang et al., 2025), OpenVLThinker-7B (Deng et al., 2025), Vision-R1-7B (Huang et al., 2025), MM-Eureka-Qwen-7B (Meng et al., 2025), ADORA-7B (Gui & Ren, 2025), ThinkLite-7B-VL (Wang et al., 2025), and VLAA-Thinker (3B/7B) (Chen et al., 2025a). We also compare vanilla GRPO (Shao et al., 2024), NoisyRollout (Liu et al., 2025), and SEED-GRPO (Chen et al., 2025b).

**Evaluation.** We assess out-of-domain generalization on five benchmarks: four reasoning (MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), MathVista (Lu et al., 2023), WeMath (Qiao et al., 2025)) and one perception (HallusionBench (Guan et al., 2024)). In our main experiments, the evaluated VLM generates the response. Qwen3-235B-A22B (Team, 2025) is used only to extract its final answer on all five OOD benchmarks. It does not generate the task response or judge its correctness; the extracted answer is scored according to the benchmark evaluation protocol. Our evaluation suite is used for Vision-R1, NoisyRollout, SEED-GRPO, the Qwen base models, and our GRPO/DSE-GRPO runs; other baseline scores are taken from the cited works.

Table 2: Full comparison of smaller VLMs on the five OOD benchmarks. Notation follows Table 1; DSE-GRPO rows are shaded.

| Model | Data Size | MVerse | MVis | MVista | WeM | Hallu | Avg. |
|---|---|---|---|---|---|---|---|
| *Open-source* | | | | | | | |
| InternVL2.5-4B | - | 37.1 | 20.9 | 60.5 | - | 46.3 | - |
| InternVL3-2B | - | 25.3 | 21.7 | 57.0 | 22.4 | 42.5 | 33.8 |
| Mulberry-2B | - | - | **51.7** | - | - | 44.9 | - |
| *R1-related (RL-tuned with verifiable reward)* | | | | | | | |
| R1-VL-2B | 260K+10K | 26.2 | 17.1 | 52.1 | - | 44.0 | - |
| R1-OneVision-3B | 155K+10K | 38.6 | 23.7 | - | - | - | - |
| VLAA-Thinker-3B | 126K+15K | 36.4 | 24.4 | 61.0 | 33.8 | - | - |
| Qwen2.5-VL-3B-Instruct | - | 33.2 | 22.7 | 58.9 | 54.0 | 60.5 | 45.9 |
| + Vanilla GRPO | 2.1K | 38.8 | 24.9 | 61.7 | 61.2 | 60.0 | 49.3 |
| + DSE-GRPO | 2.1K | 40.4 | 26.3 | 62.9 | 62.9 | 60.4 | 50.6 |
| + Vanilla GRPO | 6.4K | 42.6 | 25.5 | 62.8 | 63.6 | **61.0** | 51.1 |
| + DSE-GRPO | 6.4K | **43.6** | 26.9 | **64.2** | **64.1** | 60.1 | **51.8** |

**Implementation.** We use EasyR1 (Zheng et al., 2025b), built on verl (Sheng et al., 2024). The Qwen2.5-VL experiments follow Liu et al. (2025); Chen et al. (2025b) and run on 8×A100 with a frozen vision encoder, batch size 128, rollout batch size 512, temperature 1.0, learning rate $1\times10^{-6}$, and `token-mean` policy-loss aggregation. DSE-GRPO uses $n_1 = n_2 = 6$, $\alpha = \beta = 0.01$, and $f(x) = g(x) = 1 - x$. Appendix D lists the available training configurations.

**Semantic clustering.** Final answers are emitted inside a `\box{}` field. We form semantic clusters deterministically by grouping rollouts with identical normalized `\box{}` answers (after stripping whitespace/punctuation). This answer-based equivalence requires neither an embedding model nor a similarity threshold. Clustering and SE computation take $O(G)$ time. For the standard $G \leq 12$ setting, the measured overhead is under 1 ms per batch (Section 4.2). Two rollouts with the same answer but different reasoning chains are treated as equivalent, consistent with the verifiable reward used in GRPO (see Appendix B).

### 4.1 Main Results

**Main results across backbones.** Table 1 reports the results for Qwen2.5-VL-7B and Qwen3-VL-8B. DSE-GRPO improves the reported five-benchmark average over both base models, by 8.0 and 1.9 percentage points, respectively. The Qwen2.5-VL-7B result uses MMK12; the Qwen3-VL-8B result uses Geometry3K and MMK12 together. The gains apply to the respective backbone and training-data settings; the different training sets preclude attributing the difference in gain solely to model family. Comparisons across model families and training-data budgets are descriptive rather than matched estimates of the algorithm's effect. In the Qwen2.5-VL comparisons, vanilla GRPO uses 12 rollouts and DSE-GRPO uses six clean and six noisy rollouts.

**Comparison with alternative updates.** For Qwen2.5-VL-7B trained on MMK12, DSE-GRPO exceeds vanilla GRPO and NoisyRollout by 2.1 and 1.7 points in OOD average accuracy. With Geometry3K training, it exceeds NoisyRollout by 1.5 points and SEED-GRPO by 1.3 points (Table 1). The SEED-GRPO comparison and the coefficient ablation in Section 4.2 support the additional differential factor empirically. Bias reduction still requires the gradient-alignment condition in Analysis 1.

**Qwen3-VL-8B results.** The combined-data experiment improves all five benchmark scores over the Qwen3-VL-8B base model: MathVerse by 2.6 points, MathVision by 1.7, MathVista by 1.6, WeMath by 3.1, and HallusionBench by 0.4. The largest gain is on WeMath, and the positive HallusionBench change extends the improvement to visual perception. This base-model comparison extends the evaluation to another backbone. Comparisons against vanilla GRPO, NoisyRollout, and SEED-GRPO use Qwen2.5-VL.

**Smaller models and broader tasks.** On Qwen2.5-VL-3B, DSE-GRPO improves the OOD average over vanilla GRPO by 1.3 points with Geometry3K and 0.7 points with MMK12 (Table 2). The gains

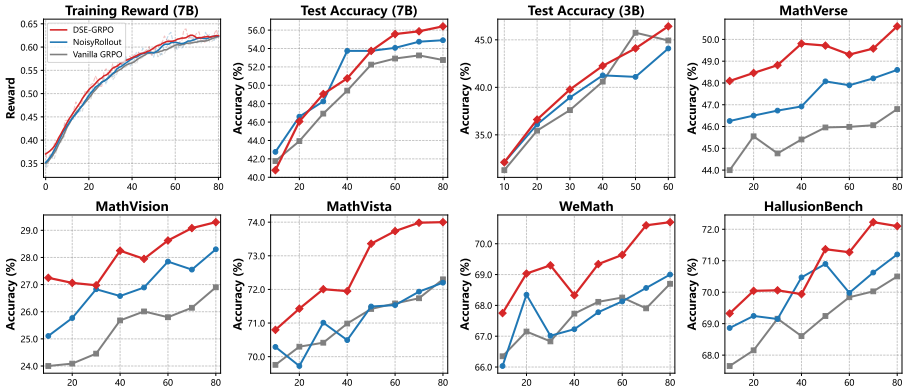



Figure 3: Training dynamics for Qwen2.5-VL. The first three panels show 7B training reward and 7B/3B in-domain test accuracy; the remaining panels show OOD accuracy. Curves compare DSE-GRPO, NoisyRollout, and vanilla GRPO at the same total rollout count. The in-domain test curves are smoother under DSE-GRPO than vanilla GRPO in the reported runs.

are not uniform across benchmarks: in the MMK12 setting, HallusionBench decreases by 0.9 points despite improvements on all four reasoning benchmarks. Appendix C reports additional ChartBench and MMMU experiments, which assess the method beyond geometry-focused training.

**Stable optimization.** The 7B and 3B test-accuracy panels in Figure 3 show smoother in-domain accuracy under DSE-GRPO than vanilla GRPO. These curves measure accuracy rather than gradient variance; Analysis 2 explains why bounded reweighting alone does not guarantee variance reduction.

**Entropy and rollout diversity.** In the reported training curves, Figure 4 shows slower policy-entropy decay and preserved rollout diversity under DSE-GRPO relative to NoisyRollout. Both methods show declining diversity after the early peak, with DSE-GRPO remaining close to NoisyRollout. The entropy curve should therefore be interpreted together with the diversity curve: higher policy entropy alone does not establish broader or more useful reasoning coverage. These observations suggest that the reweighting can coexist with continued exploration in this setting. The local bound in Analysis 3 does not guarantee this entropy ordering. Failure cases are discussed in Appendix B.

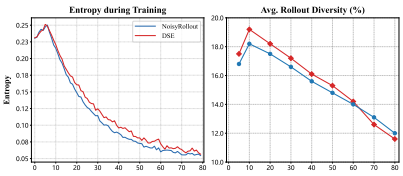



Figure 4: Policy entropy (left) and rollout diversity (right) during Qwen2.5-VL training. Relative to NoisyRollout, DSE-GRPO shows slower entropy decay and comparable rollout diversity in this run.

### 4.2 Ablation Study

**Uncertainty coefficients $\alpha$ and $\beta$.** Table 3(a,b) separates the contributions of absolute and differential entropy. With $\beta = 0$, the best tested absolute-entropy setting raises the OOD average from 57.9% to 58.7%. Adding the differential term at $\alpha = \beta = 0.01$ raises it to 59.4%, exceeding every tested $\alpha$-only configuration. Larger tested coefficients reduce the average gain, supporting moderate attenuation in this setting.

**Rollout number and noisy proportion.** Table 3(c,d) varies the total rollout count and noisy share. At a fixed one-half noisy share, increasing the total from 8 to 12 improves every reported score; increasing it to 16 or 20 yields no further gain. Exact-duplicate rollouts are removed before SE computation, but this does not ensure that additional samples introduce new answer clusters.

With the total fixed at 12, the balanced six-clean/six-noisy configuration performs best among the tested allocations. The decline at nine noisy rollouts shows that a larger noisy share is not automatically beneficial.

Table 3: Ablations on Qwen2.5-VL-7B-Instruct trained on Geometry3K (accuracy, %). Panels (a,b) vary the entropy coefficients; panels (c,d) vary total rollouts and noisy share. Geo3K reports in-domain accuracy; Avg. averages the five OOD benchmarks. Bold and underline mark the best and second-best values within each panel.

| $\alpha$ | MVerse | MVis | MVista | WeM | Hallu | Avg. |
|---|---|---|---|---|---|---|
| 0 | 48.6 | 28.3 | 72.2 | 69.0 | 71.2 | 57.9 |
| 0.01 | 48.5 | **29.2** | **72.9** | **71.3** | <u>71.5</u> | **58.7** |
| 0.02 | <u>49.9</u> | 28.5 | 72.1 | 69.4 | **71.6** | <u>58.3</u> |
| 0.05 | **50.1** | <u>28.7</u> | <u>72.4</u> | <u>69.5</u> | 70.2 | 58.2 |

(a) $\alpha$ sweep with $\beta = 0$ (absolute entropy only).

| $\beta$ | MVerse | MVis | MVista | WeM | Hallu | Avg. |
|---|---|---|---|---|---|---|
| 0.005 | <u>51.0</u> | 28.1 | <u>73.1</u> | 68.8 | 71.4 | 58.5 |
| 0.01 | 50.6 | **29.3** | **74.0** | **70.7** | **72.1** | **59.4** |
| 0.02 | 49.6 | <u>28.4</u> | **74.0** | <u>70.3</u> | 70.3 | 58.5 |
| 0.05 | **52.2** | 27.8 | 72.2 | 69.3 | <u>71.8</u> | <u>58.7</u> |

(b) $\beta$ sweep with $\alpha = 0.01$ (full DSE-GRPO).

| Rollout | Geo3K | MVerse | MVis | MVista | WeM | Hallu |
|---|---|---|---|---|---|---|
| 8 | 54.7 | 48.9 | 28.4 | 73.0 | 69.8 | 71.3 |
| 12 | **56.2** | **50.6** | **29.3** | **74.0** | **70.7** | **72.1** |
| 16 | <u>55.6</u> | <u>50.0</u> | 28.6 | 72.8 | <u>70.3</u> | <u>71.8</u> |
| 20 | 55.2 | 49.6 | <u>28.9</u> | <u>73.6</u> | 70.0 | 71.5 |

(c) Total rollout number (noisy share $6/12$).

| Prop. | Geo3K | MVerse | MVis | MVista | WeM | Hallu |
|---|---|---|---|---|---|---|
| $0/12$ | 52.8 | 46.8 | 26.9 | 72.3 | 68.7 | 70.5 |
| $3/12$ | 55.0 | 49.2 | 28.3 | <u>73.6</u> | 69.9 | 71.4 |
| $6/12$ | **56.2** | **50.6** | **29.3** | **74.0** | **70.7** | **72.1** |
| $9/12$ | <u>55.5</u> | <u>49.8</u> | <u>28.7</u> | 73.3 | <u>70.1</u> | <u>71.6</u> |

(d) Noisy proportion $n_2/(n_1+n_2)$ (total fixed at 12).

**Robustness to perturbation type.** Table 4 compares Gaussian noise, which adds pixel-level variation, Gaussian blur, which reduces fine detail, and spatial occlusion, which removes local content. Gaussian noise gives the highest accuracy on all five benchmarks, followed by blur and occlusion. Noise exceeds blur by 0.4–0.8 points per benchmark.

| Perturbation | MVerse | MVis | MVista | WeM | Hallu | Avg. |
|---|---|---|---|---|---|---|
| None (Vanilla GRPO) | 46.8 | 26.9 | 72.3 | 68.7 | 70.5 | 57.0 |
| Gaussian noise | **50.6** | **29.3** | **74.0** | **70.7** | **72.1** | **59.4** |
| Gaussian blur | 49.8 | 28.5 | 73.6 | 70.1 | 71.7 | 58.7 |
| Spatial occlusion | 48.7 | 27.9 | 72.5 | 69.3 | 70.9 | 57.0 |

Table 4: Perturbation comparison on Qwen2.5-VL-7B-Instruct trained on Geometry3K (accuracy, %).

This performance ordering does not establish a monotone relationship between entropy and semantic drift: a perturbation may induce coherent but incorrect answers without increasing entropy (Appendix B).

**Computational overhead.** The additional computation consists of deterministic `\boxed{}`-answer clustering plus $O(G)$ entropy computation. With $G \leq 12$, this takes under 1 ms per batch, compared with $\sim$1200–1650 s per training step (Table 6). End-to-end wall-clock time remains within 0.1% of NoisyRollout.

## 5 Conclusion and Limitations

DSE-GRPO uses the positive entropy increase from clean to noisy groups to regulate noisy-rollout advantages in GRPO. The main experiments show higher average OOD accuracy than the Qwen2.5-VL-7B and Qwen3-VL-8B base models; Qwen2.5-VL comparisons further show gains over alternative updates and smoother training dynamics. The analysis identifies the conditions for bias reduction and the limits of samplewise attenuation and local entropy bounds.

**Limitations.** DSE can miss coherent but incorrect shifts between clean and noisy answers, and final-answer clustering cannot detect erroneous reasoning masked by answer agreement. Neither zero DSE nor bounded weights guarantee harmless or unbiased updates. Our main experiments focus on Gaussian noise, with blur/occlusion robustness in Table 4. Appendix B discusses these failure modes. Richer distribution comparisons and process-level uncertainty signals remain future directions.

## AI USE STATEMENT

In preparing this revision, OpenAI Codex assisted with LaTeX migration, language editing, table formatting, consistency checks, and revision of mathematical claims and their proofs, including the identification of limitations and counterexamples to unconditional bias, variance, and entropy-preservation guarantees. No new training runs or measured experimental results were produced during this assistance. Separately, Qwen3-235B-A22B is used for answer extraction in the evaluation protocol described in Section 4 and Appendix D. The authors are responsible for the final text, claims, and results.

## REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 describe DSE-GRPO. Appendix A.1 states the analytical assumptions and provides derivations; Section 4 and Appendix D describe the available evaluation and training settings. Tables 1 and 2 provide the full benchmark comparisons, including the combined-data Qwen3-VL-8B experiment.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025a.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025c. URL https://arxiv.org/abs/2412.05271.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, and Yaser Yacoob. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Lujun Gui and Qingnan Ren. Training reasoning model with dynamic advantage estimation on reinforcement learning. https://github.com/ShadeCloak/ADORA, 2025. Notion Blog.

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

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

Michael Jerge, Joseph Pelczar, and Justin Downes. Perturb the thought, not the pixels: Latent-space rollout diversification for reinforcement learning of vision-language models, 2026. URL https://arxiv.org/abs/2608.21595.

Senjie Jin, Peixin Wang, Boyang Liu, Xiaoran Fan, Shuo Li, Zhiheng Xi, Jiazheng Zhang, Yuhao Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. Entropy is not enough: Unlocking effective reinforcement learning for visual reasoning via vision-anchored token selection, 2026. URL https://arxiv.org/abs/2606.03937.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv*, 2023.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, Jin Zeng, et al. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025.

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

Juhong Min, Lazar Valkov, Vitali Petsiuk, Hossein Souri, and Deen Dayal Mohan. Foveated reasoning: Stateful, action-based visual focusing for vision-language models, 2026. URL https://arxiv.org/abs/2604.21079.

Aleksei Petrenko, Ben Lipkin, Kevin Chen, Erik Wijmans, Marco Cusumano-Towner, Raja Giryes, and Philipp Krähenbühl. Entropy-preserving reinforcement learning, 2026. URL https://arxiv.org/abs/2603.11682.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, and Yifan Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *ACL*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

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

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

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

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

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.

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

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

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

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

# A Additional Assumptions and Proof Details

## A.1 Assumptions and Estimators

We fix the rollout policy parameters $\theta = \theta_{\text{old}}$ and take expectations over prompts, perturbations, and response groups. The reference $G^\star$ is a fixed gradient for the clean-input objective, not a random batch estimate. For the valid-token aggregation in Eq. (5), define the response contribution

$$u_i = \frac{A_i}{N_{\mathcal{B}}} \sum_t M_{it} \, \nabla_\theta \log \pi_\theta(o_{it} \mid h_{it})|_{\theta=\theta_{\text{old}}} . \tag{9}$$

Random response lengths and the shared normalization factor are included in $u_i$, rather than assumed independent of it. We assume a finite rollout count and $\mathbb{E}(\sum_i \|u_i\|)^2 < \infty$.

Advantages, answer clusters, and entropy weights are computed from rollout data and held fixed during differentiation. This stop-gradient convention does not imply statistical independence. At $\theta_{\text{old}}$, importance ratios equal one and clipping is locally inactive for $\epsilon > 0$. The weighted and unweighted local gradients are therefore $G_D = \sum_i w_i u_i$ and $G_0 = \sum_i u_i$. The comparison holds samples and aggregation fixed. It does not identify $G_0$ with every implementation of NoisyRollout, nor model repeated clipped updates, optimizer preconditioning, or weight decay.

**Frequency-based entropy and linear weights.** For $m \geq 1$ retained responses, $p_k = |C_k|/m$ is a probability distribution. Shannon entropy satisfies $0 \leq -\sum_k p_k \log p_k \leq \log K \leq \log m$. If the original group has size $n \geq m$ and $n > 1$, division by $\log n$ yields a value in $[0, 1]$. A singleton group has zero entropy. Removing duplicate response strings does not remove distinct strings solely because their final answers agree; these contribute separately to the answer counts.

With $f(x) = g(x) = 1 - x$ and $\alpha, \beta \in [0, 1]$, the clean and noisy weights lie in $[(1 - \alpha)(1 - \beta), 1]$. Relative to the absolute-entropy weight $w_n^{\text{abs}} = 1 - \alpha \text{SE}_n$, the additional attenuation is exactly

$$w_n^{\text{abs}} - w_n = \beta \text{DSE}(1 - \alpha \text{SE}_n) \geq 0. \tag{10}$$

At fixed $\text{SE}_n$, this difference grows linearly with DSE. This is the selectivity supplied by the differential factor; it does not guarantee that high-DSE updates are harmful or zero-DSE updates are correct.

**Bias decomposition.** Since $G_D - G_0 = \sum_i (w_i - 1)u_i$, linearity gives $\mathbb{E}G_D - G^\star = \Delta + B$. Expanding $\|\Delta + B\|^2$ proves Eq. (6). Thus non-increasing squared bias is equivalent to $2\langle \Delta, B \rangle + \|B\|^2 \leq 0$. The weight bound alone places no such restriction on $B$.

For scalar $w_i$ and vector $u_i$, define $\text{Cov}(w_i, u_i) = \mathbb{E}[(w_i - \mathbb{E}w_i)(u_i - \mathbb{E}u_i)]$. Expansion yields

$$\mathbb{E}[w_i u_i] = \mathbb{E}[w_i]\mathbb{E}[u_i] + \text{Cov}(w_i, u_i). \tag{11}$$

An entropy-dependent multiplier cannot therefore be replaced by its mean when analyzing bias. Even deterministic shrinkage can introduce bias: if an unweighted scalar estimate equals a nonzero target exactly, multiplying it by $w < 1$ makes it biased. DSE is not a density-ratio correction for the clean-input target.

## A.2 Update Perturbation, Second Moments, and Variance

Let $S = \sum_i \|u_i\|$. Since $0 \leq 1 - w_i \leq \delta$, the triangle inequality gives

$$\|G_D - G_0\| \leq \sum_i (1 - w_i)\|u_i\| \leq \delta S, \qquad \|G_D\| \leq \sum_i w_i \|u_i\| \leq S. \tag{12}$$

This proves Analysis 2 and yields $\mathbb{E}\|G_D - G_0\|^2 \leq \delta^2 \mathbb{E}S^2$ and $\|B\| \leq \delta \mathbb{E}S$. These are perturbation bounds, not statements that the update approaches the target gradient.

The ordering $\mathbb{E}\|G_D\|^2 \leq \mathbb{E}\|G_0\|^2$ does not follow. Take opposing scalar contributions $u_1 = 1$, $u_2 = -1$ and weights $w_1 = 1$, $w_2 = 1 - \varepsilon_w$, where $0 < \varepsilon_w \leq \delta$. Then $G_0 = 0$ but $G_D = \varepsilon_w$. Reweighting changes the negative cross term responsible for cancellation. The counterexample works

for weights arbitrarily close to one and disproves any moment ordering based solely on their stated interval.

Total variance is $\mathrm{Var}(G) = \mathbb{E}\|G - \mathbb{E}G\|^2 = \mathbb{E}\|G\|^2 - \|\mathbb{E}G\|^2$. A smaller second moment does not imply smaller variance when the mean changes. For example, $G_0 = 1$ has zero variance; a weight taking 1 and $1 - \varepsilon_w$ with equal probability gives $\mathrm{Var}(G_D) = \varepsilon_w^2/4 > 0$, despite a smaller second moment.

Finally, expand $G - G^\star = (G - \mathbb{E}G) + (\mathbb{E}G - G^\star)$. The expected cross term vanishes because $\mathbb{E}[G - \mathbb{E}G] = 0$, giving

$$\mathbb{E}\|G - G^\star\|^2 = \mathbb{E}\|G - \mathbb{E}G\|^2 + \|\mathbb{E}G - G^\star\|^2. \tag{13}$$

Neither component must decrease under reweighting. Training-accuracy curves alone do not measure gradient variance.

### A.3 Local Entropy-Change Bound

For a fixed distribution $\nu$ of image–question and response-prefix contexts $h$, define $H(\theta) = \mathbb{E}_{h\sim\nu}[-\sum_v \pi_\theta(v \mid h) \log \pi_\theta(v \mid h)]$. This is policy entropy, distinct from empirical final-answer entropy. Assume $H$ is differentiable and $\nabla H$ is $L_H$-Lipschitz on every segment between $\theta$ and $\theta + \eta G_D$. The reference distribution and $\theta$ are fixed inside the expectation over update randomness.

For $d = \eta G_D$, smoothness gives

$$H(\theta + d) - H(\theta) = \langle \nabla H(\theta), d\rangle + R(d), \qquad |R(d)| \le \frac{L_H}{2}\|d\|^2. \tag{14}$$

Taking expectations, applying the triangle and Cauchy–Schwarz inequalities, and retaining the remainder gives

$$\begin{aligned} |\mathbb{E}[H(\theta + d) - H(\theta)]| &\le \eta\, |\langle \nabla H(\theta), \mathbb{E}G_D\rangle| + \frac{L_H\eta^2}{2}\mathbb{E}\|G_D\|^2 \\ &\le \eta\|\nabla H(\theta)\|\mathbb{E}\|G_D\| + \frac{L_H\eta^2}{2}\mathbb{E}\|G_D\|^2. \end{aligned} \tag{15}$$

This proves Eq. (8). The same remainder bound also gives

$$\mathbb{E}[H(\theta + d) - H(\theta)] \ge \eta\langle \nabla H(\theta), \mathbb{E}G_D\rangle - \frac{L_H\eta^2}{2}\mathbb{E}\|G_D\|^2. \tag{16}$$

Hence $\langle \nabla H(\theta), \mathbb{E}G_D\rangle \ge (L_H\eta/2)\mathbb{E}\|G_D\|^2$ is sufficient for non-decreasing expected entropy. DSE does not automatically enforce this alignment. Ordering upper bounds does not order actual entropy changes, and changes in the distribution of visited contexts lie outside this fixed-context analysis. The reported entropy curves remain empirical evidence rather than consequences of an entropy-preservation theorem.

## B Discussion of Failure Cases

DSE measures the positive increase in final-answer entropy under perturbed observations. It is not a distance between clean and noisy answer distributions and does not measure reasoning-process correctness. The following cases illustrate these limits.

**High intrinsic ambiguity under clean inputs.** When clean rollouts already have high semantic entropy, DSE may remain small even if noise changes the answer distribution. The absolute-entropy factor $f$ can still attenuate uncertain groups, but does not establish whether the perturbation is harmful. The differential term may therefore attenuate noisy updates too little on intrinsically ambiguous prompts.

**Coherent but incorrect semantic drift.** Suppose all clean rollouts return the correct answer $A$, whereas all noisy rollouts return the same incorrect answer $B \neq A$. Both groups have zero semantic entropy, so $\mathrm{SE}_c = \mathrm{SE}_n = \mathrm{DSE} = 0$ despite complete disagreement between their answer distributions. The differential term therefore cannot distinguish this failure from two groups that agree on $A$. More generally, noise can shift probability mass between answers without increasing entropy. A small or zero DSE does not certify semantic alignment or a useful gradient.

**Accidental semantic alignment under noisy observations.** Clean and noisy groups can also agree, in which case little differential attenuation is appropriate for the measured answer-level uncertainty. However, this benign case and the coherent-drift case can have the same DSE. Distinguishing the two cases requires information beyond the entropy gap.

**Divergent reasoning paths with identical final answers.** Clustering uses normalized final `\box{}` answers rather than intermediate reasoning. An erroneous chain that happens to reach the same answer as a correct chain is placed in the same cluster, and its process error is invisible to DSE. Different valid reasoning paths are likewise grouped together. Final-answer agreement neither proves reasoning correctness nor justifies penalizing every difference in reasoning paths.

**Implications.** DSE is an answer-dispersion signal: it can miss harmful drift, and an increase need not imply that a rollout is harmful. The present experiments do not quantify the prevalence of these failure modes, so we make no claim that they are rare. Cross-group answer-distribution comparisons or process-level signals could complement the entropy gap, but they are outside the current method. These limitations leave the reported empirical comparisons intact while restricting conclusions about comprehensive drift detection or universal bias correction.

## C Experiments on General Scenarios

We evaluate Qwen2.5-VL-7B-Instruct on ChartBench (Xu et al., 2023) and MMMU (Yue et al., 2024) to assess settings beyond geometry-focused reasoning. For each benchmark, the compared methods use its corresponding training split. We report Acc+ on the Regular chart type of the ChartBench test split and overall accuracy on the MMMU validation split. These are separate training experiments, rather than additional OOD evaluations of the Geometry3K checkpoint.

Table 5: Performance on general visual–language benchmarks. All methods are trained on the corresponding training splits of each dataset using Qwen2.5-VL-7B-Instruct as the base model. Acc+ is reported for ChartBench, and Overall Accuracy is reported for MMMU.

| Method | ChartBench (Acc+) | MMMU (Overall Acc.) |
|---|---|---|
| Qwen2.5-VL-7B-Instruct (Base) | 39.2 | 58.6 |
| + Vanilla GRPO | 45.6 | 60.1 |
| + NoisyRollout | <u>46.1</u> | <u>60.3</u> |
| + DSE-GRPO | **46.5** | **61.0** |

Table 5 shows improvements over both alternative training methods on the two benchmarks. Relative to NoisyRollout, DSE-GRPO improves ChartBench Acc+ by 0.4 points and MMMU accuracy by 0.7 points. These results support the method in the reported chart-understanding and multimodal-reasoning settings.

## D Supplementary Implementation Details

Table 6 summarizes the reported Qwen2.5-VL training settings. The Qwen3-VL-8B experiment in Table 1 uses the combined 2.1K Geometry3K and 6.4K MMK12 training samples; the configurations below are specific to Qwen2.5-VL.

**Answer extraction during evaluation.** On MathVerse, MathVision, MathVista, WeMath, and HallusionBench, Qwen3-235B-A22B extracts the final answer from the evaluated VLM response for our main experimental results. This extraction step is separate from task-response generation and benchmark scoring, and is not part of DSE-GRPO training.

Table 6: Qwen2.5-VL training configurations. Shared settings are listed once; the lower block compares the two Qwen2.5-VL-7B-Instruct training runs.

| **Parameter** | **Shared configuration** | |
|---|---|---|
| Base model | Qwen2.5-VL-Instruct | |
| Vision encoder | Frozen | |
| Global / rollout batch size | 128 / 512 | |
| Rollout temperature | 1.0 | |
| Learning rate | $1 \times 10^{-6}$ | |
| Optimizer | AdamW | |
| Policy-loss aggregation | token-mean | |
| Image distortion | Gaussian noise | |
| CPU memory | 1 TB | |
| GPU | A100-SXM4-40/80GB | |
| **Qwen2.5-VL-7B-Instruct** | **Geometry3K (2.1K)** | **MMK12 (6.4K)** |
| Initial noise ($\alpha_0$) | 500 | 450 |
| Training episodes | 20 | 12 |
| Optimization steps ($t_{\max}$) | 80 | 144 |
| Sigmoid midpoint ($\gamma$) | 40 | 40 |
| Sigmoid steepness ($\lambda$) | 30 | 60 |
| Rollout number | $n_1 = n_2 = 6$ | $n_1 = n_2 = 6$ |
| Time per step | $\sim$1200 s | $\sim$1650 s |