Local Evidence, Global Credit: Visual Evidence-Calibrated Policy Optimization for Multimodal Reasoning

Anonymous ACL submission

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has substantially improved the reasoning capability of large vision-language models (LVLMs). Most RLVR methods for LVLMs follow an outcome-driven paradigm, using the correctness of final answer as supervision. However, in multimodal reasoning, this supervision overlooks a critical question: whether a successful reasoning trajectory is actually grounded in visual evidence. To address this limitation, we propose Visual Evidence-Calibrated Policy Optimization (VEPO), which turns visual grounding into an explicit trajectory-level credit calibration signal. VEPO first measures token-level visual responses by comparing the model's likelihood under the original and perturbed visual inputs. It then estimates a visual background for the rollout-group and retains only responses above the background as sparse visual evidence anchors to a rollout-level evidence score. This score calibrates outcome-driven optimization and guides the policy toward successful trajectories with stronger visual support. Experiments on six multimodal reasoning benchmarks show that VEPO consistently improves RLVR baselines without requiring additional annotations, demonstrating its scalability for multimodal policy optimization.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) (Zhang et al., 2025a) has emerged as a powerful paradigm for improving the reasoning ability of Large Language Models (LLMs). By optimizing sampled reasoning trajectories with verifiable outcome signals, online algorithms such as GRPO (Shao et al., 2024) and DAPO (Yu et al., 2026) have achieved strong performance in text-based mathematical and logical reasoning (Guo et al., 2025; OpenAI, 2024; Yang et al., 2025a; Team et al., 2025). Motivated by this success, recent work has extended RLVR to large

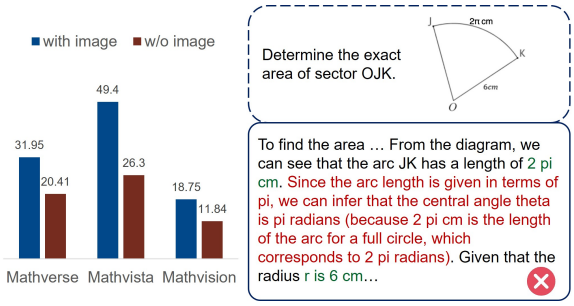


Figure 1: Analysis of reasoning on Qwen2.5-VL-3B. Left: Accuracy of with and without visual input on three multimodal reasoning datasets. Right: A case where visual tokens fail to support the final answer.

vision-language models (LVLMs), with subsequent progress driven by data-centric design (Huang et al., 2025b; Li et al., 2025a; Liang et al., 2025; Liu et al., 2026; Yang et al., 2025b), reward construction (Fan et al., 2026; Shen et al., 2025; Wang et al., 2025a; Xia et al., 2025), and other algorithmic adjustments (Zhao et al., 2026; Wang et al., 2025b; Huang et al., 2025a; Wang et al., 2026).

Although RLVR frameworks for LVLMs have achieved promising progress, outcome correctness remains insufficient to characterize whether a reasoning trajectory is actually grounded in visual perception evidence. Existing perception-aware methods highlight the value of visual perception in multimodal RLVR and mainly follow two directions: global optimization and fine-grained analysis. Global optimization methods introduce explicit rewards or implicit optimization to encourage image use (Xiao et al., 2025; Wang et al., 2025b). Although effective, such signals are usually defined at a coarse-grained level, less directly indicating critical visual evidence. Fine-grained methods further estimate token-level visual dependency and selectively optimize tokens with stronger visual responses, showing that visual dependency is unevenly distributed and often concentrates on sparse local positions (Huang et al., 2025a). This provides

a more detailed perspective, but the outcome contribution of fine-grained visual signals remain less explicit. This reveals a visual-evidence granularity mismatch in multimodal RLVR: global signals are visually under-resolved, whereas token-level signals are locally fragmented.

Our preliminary observation further supports this challenge. As shown in Figure 1, removing the image does not always lead to complete failure, suggesting that correctness alone may not reflect how much visual evidence is used. Meanwhile, token-level visual responses can reveal image dependence, but they remain fragmented across individual decoding steps. This reveals an evidence gap: the global outcome signal is visually underinformative, while local visual observations are too fragmented to serve as standalone credit. As a consequence, multimodal RLVR lacks a signal that is both trajectory-level and evidence-aware, making it difficult to calibrate complete rollouts according to their actual visual support. We view this as a weakly supervised trajectory evidence learning problem, where sparse local visual observations need to be converted into creditable trajectory-level evidence.

To this end, we propose Visual Evidence-Calibrated Policy Optimization (VEPO), a cross-granularity RLVR framework that aligns local visual evidence with trajectory-level credit. VEPO first compares generated tokens under the original and perturbed visual conditions, producing token-level responses that reveal local image dependence. It then uses the rollout group to estimate a prompt-specific response background and selects above-background responses as sparse visual evidence. These selected responses are aggregated into a trajectory-level evidence score, which is correctness-gated, group-normalized, and injected into the original advantage. In this way, VEPO enables outcome-driven optimization to account for visual support beyond final correctness.

To sum up, our main contributions are as follows:

- We propose VEPO, a cross-granularity framework that mitigates outcome ambiguity in multimodal RLVR by aligning sparse visual evidence with trajectory-level credit.
- We design group-wise upper-tail evidence distillation to extract above-background evidence anchors and form a correctness-gated trajectory signal, enabling sparse local evidence to contribute to rollout-level credit assignment.
- Experiments on six multimodal reasoning benchmarks show that VEPO consistently improves RLVR baselines without requiring additional annotations, demonstrating a scalable way to incorporate visual evidence into policy optimization.

2 Related Work

Multimodal Reasoning. Large language models have demonstrated remarkable reasoning capabilities in text-only domains (Guo et al., 2025), which has motivated recent efforts to extend this success to multimodal settings (Dong et al., 2025; Wang et al., 2024b). In particular, reinforcement learning approaches, including PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), and DAPO (Guo et al., 2025), have been actively explored to improve multimodal reasoning ability (Hu et al., 2024; Bai et al., 2025a; Wang and Peng, 2025; Yaowei Zheng, 2025). Existing studies mainly fall into data-centric methods, which focus on the construction and curation of visually grounded training data (Bai et al., 2025b; Li et al., 2025a; Liang et al., 2025; Wei et al., 2026), as well as the distillation of reasoning data (Chen et al., 2025; Huang et al., 2025b; Xu et al., 2025; Zhang et al., 2025b). Reward-oriented methods enrich the outcome supervision by incorporating grounding-related signals or additional visually informed rewards (Fan et al., 2026; Wan et al., 2026). Other approaches attempt to improve multimodal reasoning by modifying rollout strategies or introducing external visual tools (Liu et al., 2026; Wang et al., 2026; Wu et al., 2026; Zheng et al., 2025). However, these methods make visual information matter through an indirect, supervisory surrogate rather than taking visual evidence as a more integral part of the model's own reasoning dynamics.

Contrastive Visual Grounding. Visual grounding is a central challenge in LVLMs, as large vision-language models may produce plausible yet visually weakly grounded responses (Li et al., 2023; Leng et al., 2024). To probe and improve grounding, recent studies have used visual-input contrastive analysis, comparing model behaviors under original and altered visual conditions. VCD (Leng et al., 2024) formulates this contrast as an inference-time decoding mechanism, to suppress

generations driven by language priors. Vision Matters (Li et al., 2025b) extends this perturbation perspective to post-training, showing that many altered visual conditions can improve multimodal mathematical reasoning when integrated into SFT (Tong et al., 2024), DPO (Rafailov et al., 2023), and GRPO (Shao et al., 2024) pipelines. Recent multimodal RL methods further incorporate visual-input contrast into policy optimization, commonly using KL divergence between predictive distributions under original and degraded visual conditions as a proxy for visual grounding. Representative methods such as PAPO (Wang et al., 2025b) and VPPO (Huang et al., 2025a) use this signal to encourage perception-aware optimization and stronger visual dependence. Rather than focusing mainly on strengthening visual dependence during optimization, our method further promotes visual evidence as a substantive support for the reasoning process.

3 Preliminary

GRPO (Shao et al., 2024) extends Proximal Policy Optimization (PPO) (Schulman et al., 2017) by replacing the learned value function with a group-relative advantage estimate. In multimodal RLVR, a Large Vision-Language Model (LVLM) defines an autoregressive policy π_θ. Each training instance is represented as (I, q, a), where I is the visual input, q is the textual query, and a is the ground-truth answer. Given a multimodal prompt (I, q), the old policy $\pi_{\theta_{\text{old}}}$ generates a group of G responses, $\{o_i\}_{i=1}^{G}$. A binary outcome reward R_i is assigned to each complete response based solely on whether its final extracted answer matches the ground truth. The outcome reward is then normalized within the group to compute the advantage $\hat{A}_i$ for each response o_i:

$$\hat{A}_i = \frac{R_i - \text{mean}\left(\{R_k\}_{k=1}^{G}\right)}{\text{std}\left(\{R_k\}_{k=1}^{G}\right)}. \tag{1}$$

The policy π_θ is updated to maximize a clipped surrogate objective, where this uniform advantage $\hat{A}_i$ is broadcast to every timestep t:

$$\mathcal{L}^{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_i, \text{clip}\big(r_{i,t}(\theta),\, 1-\varepsilon,\, 1+\varepsilon\big)\hat{A}_i\Big)\Bigg], \tag{2}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta\big(o_{i,t} \mid I, q, o_{i,<t}\big)}{\pi_{\theta_{\text{old}}}\big(o_{i,t} \mid I, q, o_{i,<t}\big)}. \tag{3}$$

It represents the probability ratio. Since $\hat{A}_i$ is computed only from the final-answer reward and applied uniformly to the whole response, it cannot indicate whether a trajectory is visually grounded.

4 Method

Within the typical multimodal RLVR pipeline, we introduce Visual Evidence-Calibrated Policy Optimization (VEPO) that turns visual grounding into an explicit calibration signal. It estimates the visual support of each generated trajectory through a local-to-global process, from token-level visual responses to a trajectory-level evidence score, and uses this estimate to adjust the trajectory-level advantage. The overview framework is shown in Figure 2.

4.1 Visual Response Signal

To identify which generated tokens are actually supported by the visual input, we construct a token-level visual response signal using a perturbation of the input image, inspired by (Wang et al., 2025b). We define a token's visual response by comparing its log-probability under the original and perturbed image conditions, thereby measuring how much the visual evidence increases the model's confidence in the token that was actually generated.

Let I be the original image and I' be a non-informative, perturbed version. Given the current state $s_t = (q, o_{<t})$, the token visual response of the realized token o_t is defined as log-probability difference under the original and perturbed image conditions:

$$\Delta_t = \log \pi_\theta(o_t \mid s_t, I) - \log \pi_\theta(o_t \mid s_t, I') \tag{4}$$

A high Δ_t value indicates that the generated token becomes substantially less likely when informative visual evidence is removed, suggesting that its generation is more strongly supported by the original image. Therefore, Δ_t highlights decoding steps where the current reasoning path relies more on visual evidence rather than language context alone.

4.2 Group-wise Sparse Aggregation

Since token-level visual responses are typically sparsely distributed and exhibit a highly skewed distribution (Huang et al., 2025a), we do not treat all

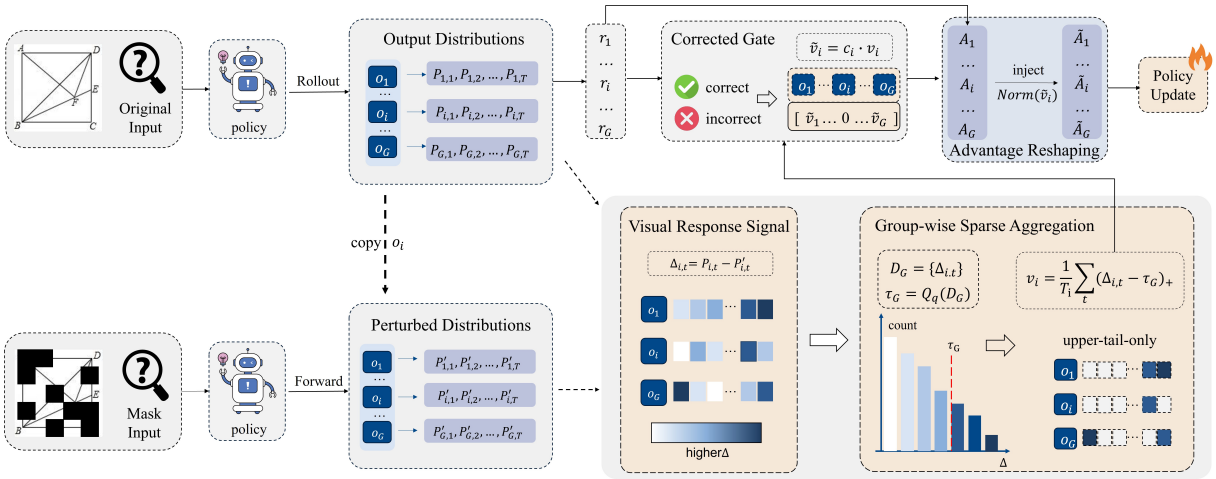


Figure 2: Overview of our VEPO framework. Given an original multimodal input, we first calculate visual response signals in token-level. The resulting token-level responses are then processed by group-wise sparse aggregation, which retains only above-background visual responses and summarizes them into a trajectory-level visual support score. Finally, VEPO applies correctness gating and injects the normalized visual support score into the original advantage, so that the policy update remains guided by verifiable rewards while assigning stronger updates to correct trajectories with stronger visual support.

tokens equally. For each prompt, let $\mathcal{G} = \{o_i\}_{i=1}^{G}$ denote the group of sampled trajectories. We collect all token-level visual evidence scores within the group:

$$\mathcal{D}_{\mathcal{G}} = \{\Delta_{i,t} \mid 1 \le i \le G,\ 1 \le t \le T_i\}. \tag{5}$$

Then, we compute a group-wise quantile threshold:

$$\tau_{\mathcal{G}} = Q_q(\mathcal{D}_{\mathcal{G}}), \tag{6}$$

where $Q_q(\cdot)$ denotes the q-th quantile. This threshold defines a prompt-specific visual background level. It identifies which token responses are unusually high compared with the visual response distribution of the same rollout group.

For each trajectory o_i, we aggregate only the excessive positive part above the threshold:

$$v_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \max(\Delta_{i,t} - \tau_G, 0). \tag{7}$$

The resulting score measures the average cumulative strength of visual evidence that stands out above the group background level along trajectory i.

4.3 Correctness-gated Advantage Calibration

Not all visually sensitive trajectories are desirable. A trajectory may respond strongly to the image and still produce an incorrect answer. Therefore, VEPO uses the verifiable outcome reward as a correctness gate, so that only successful trajectories can receive positive visual calibration.

$$\tilde{v}_i = c_i \cdot v_i, \tag{8}$$

where $c_i \in \{0, 1\}$ denotes whether the trajectory produces the correct final answer.

To make visual scores compatible with group-relative optimization, we normalize the gated score within each group:

$$b_i = \frac{\tilde{v}_i - \tilde{\mu}_G}{\tilde{\sigma}_G + \epsilon}, \tag{9}$$

where

$$\tilde{\mu}_G = \frac{1}{G} \sum_{j=1}^{G} \tilde{v}_j, \qquad \tilde{\sigma}_G = \sqrt{\frac{1}{G} \sum_{j=1}^{G} (\tilde{v}_j - \tilde{\mu}_G)^2}. \tag{10}$$

b_i is a relative preference visual signal where indicates which trajectories among the rollouts generated for the same prompt, are more likely to be correct and visually grounded.

For the i-th trajectory, the advantage is defined as:

$$\tilde{A}_i = \hat{A}_i(1 + \lambda b_i), \tag{11}$$

where λ controls the strength of the visual shaping signal.

By substituting $\tilde{A}_i$ into the GRPO surrogate objective, we derive the final policy gradient formula-

tion:

$$\mathcal{L}^{\mathrm{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\tilde{A}_i, \\ \mathrm{clip}\big(r_{i,t}(\theta),\, 1-\varepsilon,\, 1+\varepsilon\big)\tilde{A}_i\Big)\Bigg]. \tag{12}$$

This adaptively amplifies updates for trajectories with high visual engagement and dampens those that are less visually grounded.

4.4 Theoretical Analysis

We further provide a justification of the proposed sparse aggregation and its role in policy optimization. Specifically, we first show that the trajectory-level visual signal captures both the coverage and strength of above-background visual responses. We then analyze how this group-relative signal modulates the policy-gradient update in a scale-controlled manner.

Sparse visual-support assumption. For multimodal reasoning trajectories sampled from the same query, most tokens are dominated by language priors, formatting patterns, or intermediate reasoning context, and thus only exhibit background-level visual responses. Formally, the token-level response distribution of trajectory i can be viewed as a mixture:

$$\Delta_{i,t} \sim (1-\alpha_i)F_0 + \alpha_i F_i^{+}, \tag{13}$$

where F_0 denotes the group-level background response distribution, F_i^{+} denotes the visual-support component of trajectory i, and α_i denotes the proportion of visually supported tokens in this trajectory.

Based on this assumption, directly averaging all token-level responses may dilute the sparse but informative visual-support signals. Therefore, we use the group-wise threshold τ_G to approximate the background response level and retain only the exceedance above this shared baseline:

$$v_i = \frac{1}{T_i}\sum_{t=1}^{T_i}(\Delta_{i,t} - \tau_G)_+. \tag{14}$$

This makes v_i an upper-tail statistic rather than a simple average over all tokens.

Proposition 1. Let

$$A_i(\tau_G) = \{t : \Delta_{i,t} > \tau_G\} \tag{15}$$

be the set of above-threshold visual-response tokens in trajectory i. If $A_i(\tau_G) \neq \emptyset$, then

$$v_i = \frac{|A_i(\tau_G)|}{T_i} \cdot \frac{1}{|A_i(\tau_G)|}\sum_{t\in A_i(\tau_G)}(\Delta_{i,t} - \tau_G). \tag{16}$$

When $A_i(\tau_G) = \emptyset$, we define $v_i = 0$.

Proof. Since $(\Delta_{i,t} - \tau_G)_+ = 0$ for all tokens satisfying $\Delta_{i,t} \leq \tau_G$, we have

$$v_i = \frac{1}{T_i}\sum_{t\in A_i(\tau_G)}(\Delta_{i,t} - \tau_G). \tag{17}$$

Multiplying and dividing by $|A_i(\tau_G)|$ gives the result.

Proposition 1 shows that v_i jointly captures two aspects of sparse visual support. The first term,

$$\frac{|A_i(\tau_G)|}{T_i}, \tag{18}$$

measures the coverage of above-background visual-response tokens. The second term,

$$\frac{1}{|A_i(\tau_G)|}\sum_{t\in A_i(\tau_G)}(\Delta_{i,t} - \tau_G), \tag{19}$$

measures their average exceedance strength. Hence, a trajectory obtains a large v_i only when it contains sufficiently frequent or sufficiently strong visually responsive tokens.

We now analyze how the group-relative visual-support signal affects the RL update. For clarity, we ignore clipping and treat b_i as a detached rollout-level signal. Let

$$u_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\nabla_\theta \log \pi_\theta(o_{i,t} \mid I, q, o_{i,<t}) \tag{20}$$

denote the trajectory-level score-function gradient. The standard GRPO update can be written as

$$g_{\mathrm{GRPO}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i u_i\right]. \tag{21}$$

Our method uses the shaped advantage

$$\tilde{A}_i = \hat{A}_i(1 + \lambda b_i), \tag{22}$$

which yields

$$g_{\mathrm{VEPO}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i(1+\lambda b_i)u_i\right]. \tag{23}$$

Expanding the estimator gives

$$g_{\text{VEPO}} = g_{\text{GRPO}} + \lambda g_{\text{vis}}, \tag{24}$$

where

$$g_{\text{vis}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i b_i u_i\right]. \tag{25}$$

This decomposition shows that our method introduces an auxiliary visual support correction term that modulates the update strength according to the interaction between outcome advantage and group-relative visual support.

Proposition 2: Scale-controlled visual correction. Since b_i is normalized within each rollout group, it satisfies

$$\frac{1}{G}\sum_{i=1}^{G} b_i = 0, \qquad \frac{1}{G}\sum_{i=1}^{G} b_i^2 \le 1. \tag{26}$$

For each rollout group, define the visual correction term

$$c_G = \frac{1}{G}\sum_{i=1}^{G}\hat{A}_i b_i u_i. \tag{27}$$

Then

$$\begin{aligned}\|c_G\| &\le \left(\frac{1}{G}\sum_{i=1}^{G} b_i^2\right)^{1/2}\left(\frac{1}{G}\sum_{i=1}^{G}\|\hat{A}_i u_i\|^2\right)^{1/2} \\ &\le \left(\frac{1}{G}\sum_{i=1}^{G}\|\hat{A}_i u_i\|^2\right)^{1/2}.\end{aligned} \tag{28}$$

Therefore, the deviation from the original GRPO update is bounded by

$$\begin{aligned}&\|g_{\text{VEPO}} - g_{\text{GRPO}}\| = \lambda\|g_{\text{vis}}\| \\ &\le \lambda\mathbb{E}\left[\left(\frac{1}{G}\sum_{i=1}^{G}\|\hat{A}_i u_i\|^2\right)^{1/2}\right].\end{aligned} \tag{29}$$

Proof. The first equality follows from the policy-gradient decomposition above. For each rollout group, applying the Cauchy–Schwarz inequality over the trajectory index i gives

$$\begin{aligned}&\left\|\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i b_i u_i\right\| \le \frac{1}{G}\sum_{i=1}^{G}|b_i|\cdot\|\hat{A}_i u_i\| \\ &\le \left(\frac{1}{G}\sum_{i=1}^{G} b_i^2\right)^{1/2}\left(\frac{1}{G}\sum_{i=1}^{G}\|\hat{A}_i u_i\|^2\right)^{1/2}.\end{aligned} \tag{30}$$

This result shows that our method retains the original outcome-driven GRPO update as the leading term, while adding a scale-controlled visual-support correction. Since b_i is normalized within each rollout group, the magnitude of the correction term is bounded by the root-mean-square magnitude of the original advantage-weighted trajectory gradients.

5 Experiment

5.1 Experimental Setup

Implementation Details. To balance training efficiency and data diversity, we construct a mixed training subset from ViRL39K (Wang et al., 2026) and Geometry3K (Lu et al., 2021). Specifically, we randomly sample 13K instances from ViRL39K(39K) and 0.7K instances from Geometry3K(21K), and convert them into a unified RLVR training format. All methods are trained on the same mixed subset for 2 epochs with a learning rate of 1e-6. We apply Qwen2.5-VL-3B as the base model, comparing the standard GRPO and DAPO baselines with our proposed method. Following previous work, we generate 5 responses per example with a sampling temperature of 1.0. We set the maximum response length to 2048 for all models. For our method, we set group-wise quantile ratio to $q = 0.6$ and the visual shaping signal to $\lambda = 0.1$. Additional implementation details are provided in Appendix A.

Benchmarks. We evaluate multimodal understanding and reasoning on six diverse benchmarks. The benchmarks cover math, logic, geometric, and multi-discipline multimodal reasoning, including MathVerse (Zhang et al., 2024), MMK12 (Meng et al., 2025), MathVision (Wang et al., 2024a), MathVista (Lu et al., 2023), LogicVista (Xiao et al., 2024), MMMU-Pro (Yue et al., 2025). All results are based on exact match against ground truth answer, avoiding reliance on LLM-as-a-judge systems. We report accuracy using a fixed greedy decoding configuration with temperature 0.

5.2 Main Results

As shown in Table 1, VEPO achieves the best performance across all evaluated multimodal reasoning benchmarks, outperforming the base model as well as the GRPO and DAPO baselines. Compared with DAPO, VEPO further improves the average performance by 1.61% and consistently brings gains on all benchmarks. These results

Method	MMMU-Pro	MathVista	MathVerse	LogicVista	MathVision	MMK12	Avg.
Qwen2.5-VL-3B	27.28	49.40	31.95	31.25	18.75	40.44	33.18
GRPO	29.88	62.20	37.44	40.40	20.72	50.00	40.11
DAPO	29.42	63.20	40.25	41.96	22.04	51.05	41.32
VEPO	**30.93**	**63.60**	**42.39**	**44.20**	**23.36**	**53.10**	**42.93**

Table 1: Comparison of different methods on multimodal reasoning benchmarks. Scores are reported as exact-match accuracy (%) on verifiable instances for objective results, avoiding any LLM-as-a-judge, and the best results are highlighted in bold.

Method	MMMU-Pro	LogicVista	MMK12	Avg.
DAPO	29.42	41.96	51.05	40.81
VEPO-Avg	30.17	40.40	**53.10**	41.22
VEPO	**30.93**	**44.20**	**53.10**	**42.74**

Table 2: Ablation on visual evidence aggregation strategy.

Method	MMMU-Pro	LogicVista	MMK12	Avg.
DAPO	29.42	41.96	51.05	40.81
VEPO-Add	28.77	40.40	**53.85**	41.01
VEPO	**30.93**	**44.20**	53.10	**42.74**

Table 3: Ablation on advantage calibration strategy.

demonstrate that VEPO provides a general and effective enhancement to multimodal RLVR by complementing outcome correctness with explicit visual-support calibration.

5.3 Analysis

Ablation Study on Visual Evidence Aggregation. We examine whether the proposed sparse aggregation is necessary for constructing the trajectory-level visual evidence signal. We compare VEPO with an average-pooling variant, denoted as VEPO-Avg, which replaces the upper-tail sparse aggregation with a direct average over all token-level visual responses:

$$v_i^{\text{avg}} = \frac{1}{T_i} \sum_{t=1}^{T_i} \Delta_{i,t}. \tag{31}$$

All other components are kept unchanged.

As shown in Table 2, VEPO-Avg slightly improves over DAPO, indicating that token-level visual responses provide useful complementary information for outcome-driven optimization. However, it remains weaker than the proposed VEPO, suggesting that how these responses are aggregated is critical. Since $\Delta_{i,t}$ is signed, dense averaging may mix positive visual evidence with noisy or negative responses: $\Delta_{i,t} > 0$ indicates stronger support from the original image, while $\Delta_{i,t} < 0$ means the token becomes more likely under the perturbed visual condition, which may reflect language priors, task-irrelevant fluctuations, or non-visual tokens. As a result, averaging all tokens can dilute sparse positive evidence, making dense averaging less reliable for trajectory-level visual support estimation.

In contrast, VEPO aggregates only above-background responses within the rollout group, focusing on tokens with relatively stronger support from the original image. By reducing the influence of language-dominated or negative responses, VEPO better captures the sparse nature of visual evidence and achieves more reliable trajectory-level evidence estimation.

Ablation Study on Advantage Calibration Strategy. We conduct an ablation study to examine how trajectory-level visual evidence should be incorporated into policy optimization. We compare the proposed multiplicative advantage calibration with an additive variant, VEPO-Add, where

$$\tilde{A}_i = \hat{A}_i + \lambda b_i, \tag{32}$$

while all other components remain unchanged. As shown in Table 3, VEPO-Add only brings marginal gains over DAPO and fails to improve consistently across benchmarks, suggesting that adding visual evidence as an independent advantage term may partially decouple the update from the outcome objective. In contrast, VEPO preserves the direction of the original outcome-driven advantage and uses visual evidence to adaptively adjust its strength. These results show that visual evidence is more effective as a calibration factor than as an independent additive signal.

Effect of the Quantile Threshold. We study the effect of the group-wise quantile ratio q, which con-

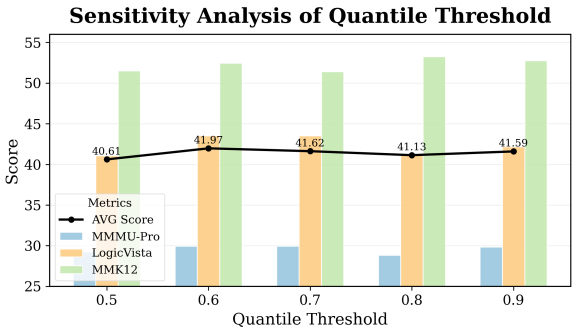


Figure 3: Sensitivity analysis of the quantile threshold under a fixed calibration weight $\lambda = 0.2$. The line shows the average score, while bars show per-benchmark score.

λ	MMMU-Pro	LogicVista	MMK12	Avg.
0.1	**30.93**	44.20	**53.10**	**42.74**
0.2	29.94	43.53	52.45	41.97
0.3	29.31	**45.54**	51.20	42.02

Table 4: Effect of the visual shaping strength λ under a fixed threshold $q = 0.6$.

trols the sparsity of the selected visual response signal. To reduce computational cost, we vary q under a fixed representative calibration weight $\lambda = 0.2$. As shown in Figure 3, VEPO is moderately affected by the threshold choice, but the overall performance remains relatively stable across a range of thresholds. A lower threshold retains more token-level responses and may introduce background visual sensitivity or language-context noise, while an overly high threshold makes the evidence signal too sparse and may discard useful visual evidence. Among the tested settings, $q = 0.6$ achieves the best average performance, suggesting that moderately sparse upper-tail selection provides a better balance between filtering noisy local responses and preserving informative visual evidence. These results indicate that VEPO benefits from sparse evidence selection while not being overly sensitive to the exact threshold value.

Effect of the Visual Shaping Strength. We further analyze the effect of the visual shaping coefficient λ, which controls the strength of the auxiliary visual evidence signal in policy optimization. As shown in Table 4, $\lambda = 0.1$ achieves the best average performance among the tested settings. Increasing λ to 0.2 or 0.3 does not further improve the average result, although $\lambda = 0.3$ obtains the highest score on LogicVista. This suggests that a relatively small visual shaping strength is sufficient to guide the policy toward visually supported trajectories, while larger weights may overemphasize the auxiliary evidence signal and weaken the balance with the original outcome-based advantage. Therefore, we use $\lambda = 0.1$ as the default setting in our main experiments. These results indicate that VEPO benefits from a lightweight visual calibration signal, rather than relying on a large auxiliary shaping weight.

Method	ScienceQA	TextVQA	Avg.
Qwen2.5-VL-3B	81.06	78.73	79.90
GRPO	81.16	79.33	80.25
DAPO	81.61	79.33	80.47
VEPO	**81.80**	**79.71**	**80.71**

Table 5: Performance on out-of-domain VQA benchmarks.

General Visual-Language Capability. To examine whether VEPO affects general visual-language capabilities, we evaluate the trained models on two out-of-domain VQA datasets, ScienceQA (Lu et al., 2022) and TextVQA (Singh et al., 2019). As shown in Table 5, VEPO maintains comparable or slightly better performance than the RLVR baselines on both out-of-domain VQA benchmarks. This suggests that visual evidence calibration does not sacrifice general VQA ability while improving multimodal reasoning. These results further indicate that VEPO enhances reasoning-oriented policy optimization without impairing the model's general visual-language capability.

6 Conclusion

In this work, we revisit multimodal RLVR from the perspective of evidence-aware trajectory credit. We identify a granularity mismatch: final-answer rewards supervise complete rollouts, whereas visual evidence often appears as sparse local responses. To bridge this gap, we propose VEPO, which distills above-background visual responses into rollout-level evidence credit to calibrate the original advantage. Experiments on multiple multimodal reasoning benchmarks show consistent improvements over GRPO and DAPO without external supervision, highlighting the value of visual-evidence calibration beyond answer correctness.

Limitations

Although VEPO improves outcome-driven multimodal policy optimization, it still has several limitations that could be recognized in future research.

Due to computational resource constraints, our experiments are conducted with limited model scales and training data. We are unable to fully verify the scalability of VEPO on larger LVLMs, such as 7B and 32B models. In addition, our evaluation mainly focuses on multimodal reasoning tasks with verifiable answers. Future work could extend VEPO to more open-ended multimodal tasks, such as image-grounded dialogue and detailed image captioning, where visual faithfulness, response quality are more difficult to verify automatically.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. 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. 2025b. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

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.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Wang. 2026. Grit: Teaching mllms to think with images. *Advances in Neural Information Processing Systems*, 38:116522–116543.

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

Jian Hu, Xibin Wu, Zilin Zhu, Weixun Wang, Dehao Zhang, Yu Cao, and 1 others. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 6.

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

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

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*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Yuting Li, Lai Wei, Kaipeng Zheng, Jingyuan Huang, Linghe Kong, Lichao Sun, and Weiran Huang. 2025b. Vision matters: Simple visual perturbations can boost multimodal math reasoning. *arXiv e-prints*, pages arXiv–2506.

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*.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Shieh. 2026. Noisyrollout: Reinforcing visual reasoning with data augmentation. *Advances in Neural Information Processing Systems*, 38:2923–2957.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *CoRR, abs/2310.02255*, 5:15.

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. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6774–6786.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, and 1 others. 2025. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

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. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, C Du, C Liao, and 1 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms, 2025. *URL https://arxiv. org/abs/2501.12599*, 118.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Advances in Neural Information Processing Systems*, 37:7821–7846.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, and 1 others. 2026. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *Advances in Neural Information Processing Systems*, 38:153676–153713.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2026. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891.

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. 2025a. 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*.

Xiaodong Wang and Peixi Peng. 2025. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video.

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

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, Jia Wang, and 1 others. 2026. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *Advances in Neural Information Processing Systems*, 38:4463–4494.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, ChengXiang Zhai, and Klara Nahrstedt. 2026. VTool-R1: VLMs Learn to Think with Images via Reinforcement Learning on Multimodal Tool Use. In *International Conference on Learning Representations*.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon 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. Perception-r1: Advancing multimodal reasoning capabilities of mllms 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*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2376–2385.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, and 1 others. 2025a. 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.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025b. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662.

Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2026. Absolute zero: Reinforced self-play reasoning with zero data. *Advances in Neural Information Processing Systems*, 38:105816–105879.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

A Implementation Details

Training Details. Our implementation is built upon the EasyR1 framework (Yaowei Zheng, 2025; Sheng et al., 2025). All experiments were conducted using PyTorch 2.6.0 with CUDA 12.4. The base model for our experiments is the open-source Qwen2.5-VL-3B. We train all models for two epochs on a mixed training subset from ViRL39K (Wang et al., 2026) and Geometry3K (Lu et al., 2021). Specifically, we randomly sample 13K instances from ViRL39K(39K) and 0.7K instances from Geometry3K(21K), and convert them into a unified RLVR training format. The vision tower is unfrozen during training. For the online RL process, we generate 5 responses per question. The final reward is a weighted sum of accuracy and format compliance, defined as $R = 0.95S_{\text{acc}} + 0.05S_{\text{format}}$, 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 training objective follows the DAPO recipe, incorporating dynamic sampling, clip-higher, and a token-level policy gradient loss, without a KL divergence penalty. All key hyperparameters for the optimizer, RL process, and evaluation are detailed in Table 6.

VEPO Configuration. Visual response estimation is an important component of VEPO, as it provides the counterfactual signal used to measure whether a generated trajectory is supported by the

Hyperparameter	Value
General Training	
Optimizer	AdamW
Learning Rate	1×10^{-6}
LR Schedule	Constant (no warmup or decay)
Epochs	2
Freeze Vision Tower	False
RL Process	
Global Batch Size	128
Rollout Batch Size	384
Rollouts per Prompt	5
Rollout Top-p	0.99
Max Response Length	2048
DAPO Recipe	
Sampling Method	Dynamic Sampling
Clip Ratio Low	0.2
Clip Ratio High	0.28
Loss Averaging Mode	Token-level
KL Penalty	None
VEPO Specific	
Quantile ratio τ	0.6
Visual signal λ	0.1
Evaluation Generation	
Temperature	0.0
Max New Tokens	4096

Table 6: Key hyperparameters for training and evaluation.

visual input. Following PAPO (Xiao et al., 2025), we compare the log-probabilities of the generated tokens under the original image I and a perturbed image I'. For the perturbation strategy, we adopt Random Patch Blackening: the image is divided into non-overlapping patches of size 14×14, and each patch is independently set to black with a probability of 0.6. This produces a degraded visual condition while keeping the input format unchanged.

Computational Resource. All models were trained on a cluster of 2 x NVIDIA A100 80GB GPUs.

B Theoretical Proof of Visual Responses

In this section, we provide a theoretical justification for the visual response signal used in VEPO. The goal of this analysis is to clarify why it is a natural choice for estimating the visual support of a sampled trajectory.

Counterfactual visual conditions. Given a multimodal prompt consisting of an image I and a textual query q, let I' denote a perturbed version of the image where informative visual content is partially removed. At decoding step t, the model has generated a prefix $o_{<t}$, and the current decoding state is denoted as

$$s_t = (q, o_{<t}). \tag{33}$$

We define the next-token distributions under the original and perturbed visual conditions as

$$p_t(\cdot) = \pi_\theta(\cdot \mid s_t, I), \quad \tilde{p}_t(\cdot) = \pi_\theta(\cdot \mid s_t, I'). \tag{34}$$

Since the textual state s_t is kept fixed, the difference between these two distributions reflects the effect of changing the visual input.

Generated-token visual response. For the token o_t actually generated in the sampled rollout, VEPO defines its visual response as

$$\Delta_t = \log \pi_\theta(o_t \mid s_t, I) - \log \pi_\theta(o_t \mid s_t, I'), \tag{35}$$

therefore,

$$\Delta_t = \log \frac{p_t(o_t)}{\tilde{p}_t(o_t)}. \tag{36}$$

This quantity is a generated-token log-likelihood ratio between the original and perturbed visual conditions. If $\Delta_t > 0$, the generated token becomes more likely when the original image is available, suggesting that this token is positively supported by visual evidence. If $\Delta_t \approx 0$, the token is weakly affected by the visual input. If $\Delta_t < 0$, the generated token is not positively supported by the original image under this counterfactual comparison.

This signed property is important for our purpose. VEPO aims to estimate whether the sampled trajectory itself is visually supported, rather than merely whether the decoding state is visually sensitive. Therefore, we use the log-ratio of the actually generated token as the basic unit of visual response.

Trajectory-level decomposition. For an autoregressive model, the probability of a complete response $o = (o_1, \ldots, o_T)$ under the original image can be factorized as

$$\pi_\theta(o \mid q, I) = \prod_{t=1}^{T} \pi_\theta(o_t \mid s_t, I). \tag{37}$$

Similarly, under the perturbed image I', we have

$$\pi_\theta(o \mid q, I') = \prod_{t=1}^{T} \pi_\theta(o_t \mid s_t, I'). \tag{38}$$

Therefore, the sequence-level log-likelihood ratio can be decomposed into token-level visual responses:

$$\log \frac{\pi_\theta(o \mid q, I)}{\pi_\theta(o \mid q, I')} = \sum_{t=1}^{T} \log \frac{\pi_\theta(o_t \mid s_t, I)}{\pi_\theta(o_t \mid s_t, I')} = \sum_{t=1}^{T} \Delta_t. \tag{39}$$

This shows that the proposed token-level response is not an isolated heuristic. Instead, it provides a natural additive decomposition of how much the sampled response is more supported under the original visual condition than under the perturbed visual condition.

Relation to distributional KL. The proposed visual response is also closely related to distributional KL divergence. At decoding step t, the KL divergence between the original-image distribution and the perturbed-image distribution is

$$D_{\mathrm{KL}}(p_t \| \tilde{p}_t) = \mathbb{E}_{y \sim p_t}\left[\log \frac{p_t(y)}{\tilde{p}_t(y)}\right]. \tag{40}$$

Since

$$\Delta_t = \log \frac{p_t(o_t)}{\tilde{p}_t(o_t)}, \tag{41}$$

we have

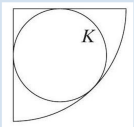
$$\mathbb{E}_{o_t \sim p_t}[\Delta_t] = D_{\mathrm{KL}}(p_t \| \tilde{p}_t). \tag{42}$$

Thus, Δ_t can be viewed as a generated-token counterpart of the log-ratio whose expectation recovers the distributional KL divergence.

This relation indicates that Δ_t is connected to the same information-gain perspective as KL-based visual dependency measures. However, the two quantities emphasize different levels of visual dependence. Distributional KL measures how much the visual input changes the entire predictive distribution at a decoding state. In contrast, Δ_t measures whether the token actually generated in the sampled rollout becomes more likely under the original image than under the perturbed image. Therefore, KL is suitable for identifying visually sensitive decoding positions, while the generated-token log-ratio is directly aligned with estimating visual support for the sampled trajectory.

KL and Δ_t capture related but different notions of visual dependence. Prior KL-based methods often use distributional visual dependency to locate visually important positions or to guide token-level gradient filtering. VEPO instead focuses on trajectory-level advantage calibration, where the central question is whether the actually sampled response is supported by visual evidence. For this purpose, the generated-token form provides a direct proxy for visual support of the generated trajectory.

Case1 generated by DAPO:



Question: A circle K is inscribed in a quarter circle with radius 6 as shown in the figure. What is the radius of circle K?
Choices: A. $\frac{6-\sqrt{2}}{2}$ B. $\frac{3\sqrt{2}}{2}$ C. 2.5 D. 3 E. $6(\sqrt{2}-1)$
Ground truth: E. $6(\sqrt{2}-1)$

<think>
To find the radius of the inscribed circle K in a quarter circle with radius 6, we can use geometric properties and relationships. Let's denote the radius of the inscribed circle K as r. First, consider the right triangle formed by the radius of the quarter circle, the radius of the inscribed circle, and the segment of the radius of the quarter circle that is not covered by the inscribed circle. The right triangle has one leg equal to the radius of the quarter circle minus the radius of the inscribed circle, and the other leg equal to the radius of the inscribed circle. The hypotenuse of this right triangle is the radius of the quarter circle.
The radius of the quarter circle is 6, so the right triangle has legs of lengths $6-r$ and r, and a hypotenuse of length r. Using the Pythagorean theorem, we get:

$$(6-r)^2+r^2 = 6^2$$

Expanding and simplifying the equation:

$$(6-r)^2+r^2 = 6^2$$
$$36-12r+r^2+r^2 = 36$$
$$2r^2-12r = 0$$

Factoring out $2r$:

$$2r(r-6) = 0$$

This gives us two solutions: $r = 0$ (which is not possible for a circle) and $r = 3$.
However, we need to verify this solution by checking the geometric configuration. The correct approach involves recognizing that the radius of the inscribed circle in a quarter circle is given by the formula:

$$r = \frac{6}{1+\sqrt{2}}$$

(Despite misinterpreting the image, the model suddenly arrives at the correct geometric formula without any prior reasoning.)
Simplifying this expression:

$$r = \frac{6}{1+\sqrt{2}} \cdot \frac{1-\sqrt{2}}{1-\sqrt{2}} = \frac{6(1-\sqrt{2})}{1-2} = \frac{6(1-\sqrt{2})}{-1} = 6(\sqrt{2}-1)$$

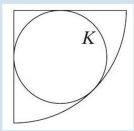
Thus, the radius of the inscribed circle is $6(\sqrt{2}-1)$.
<think>
Therefore, the correct answer is:
$\boxed{E}$

Figure 4: Case study 1 for DAPO: Correct answer with unreliable intermediate reasoning.

C Case Study

To provide a qualitative understanding of VEPO, we present two representative case studies, as shown in Figures 4–6. These examples show that outcome-only supervision can provide incomplete guidance for multimodal reasoning: a correct final answer may still be produced through unreliable intermediate reasoning, while an incorrect answer may result from misinterpreting key visual evidence. By incorporating trajectory-level visual evidence calibration, VEPO complements outcome rewards with an explicit estimate of visual support. These cases qualitatively show that VEPO can guide policy optimization toward reasoning trajectories that are not only more outcome-correct, but also better grounded in the visual input.

Case1 generated by VEPO:



Question: A circle K is inscribed in a quarter circle with radius 6 as shown in the figure. What is the radius of circle K?
Choices: A. $\frac{6-\sqrt{2}}{2}$ B. $\frac{3\sqrt{2}}{2}$ C. 2.5 D. 3 E. $6(\sqrt{2}-1)$
Ground truth: E. $6(\sqrt{2}-1)$

<think>
To find the radius of the inscribed circle K in the quarter circle with radius 6, we need to analyze the geometric configuration of the circle and the quarter circle.
1. **Identify the key elements:**
- The radius of the quarter circle is 6.
- The circle K is inscribed in the quarter circle, meaning it is tangent to the two radii of the quarter circle and the arc of the quarter circle.
2. **Set up the coordinate system**:
- Place the center of the quarter circle at the origin $(0,0)$.
- The quarter circle extends from the positive x-axis to the positive y-axis, with a radius of 6.
3. **Determine the radius of the inscribed circle:**
- Let the radius of the inscribed circle K be r.
- The center of the inscribed circle K will be at a distance r from both the x-axis and the y-axis, so its coordinates will be (r,r).
4. **Use the distance from the center of the quarter circle to the center of the inscribed circle:**
- The distance from the center of the quarter circle $(0,0)$ to the center of the inscribed circle (r,r) is $r\sqrt{2}$ (since the center of the inscribed circle is at a distance r from both axes, forming a right triangle with legs of length r .
- This distance must be equal to the radius of the quarter circle minus the radius of the inscribed circle, because the inscribed circle is tangent to the arc of the quarter circle.
Therefore, we have:

$$r\sqrt{2} = 6 - r$$

5. **Solve for r :**
- Rearrange the equation to isolate r :

$$r\sqrt{2} + r = 6$$
$$r(\sqrt{2}+1) = 6$$
$$r = \frac{6}{\sqrt{2}+1}$$

- To rationalize the denominator, multiply the numerator and the denominator by the conjugate of the denominator:

$$r = \frac{6}{\sqrt{2}+1} \cdot \frac{\sqrt{2}-1}{\sqrt{2}-1} = \frac{6(\sqrt{2}-1)}{(\sqrt{2})^2-(1)^2} = \frac{6(\sqrt{2}-1)}{2\text{-}1} = 6(\sqrt{2}-1)$$

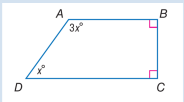
<think>
Therefore, the radius of the inscribed circle K is $\boxed{E}$.

Figure 5: Case study 1 for VEPO: Correct answer with more visually grounded reasoning.

D LLM Usage Statement

During the preparation of this paper, we used a large language model (LLM) as an assistive tool. Its use was limited to proofreading grammatical and spelling errors, as well as rephrasing sentences to improve readability and clarity. The LLM was not used to generate the core ideas, conduct data analysis, or produce the main scientific content of the paper. All intellectual contributions, research decisions, and the final manuscript remain the sole responsibility of the authors.

Case2:



Question: Find $m\angle A$ of quadrilateral $ABCD$.
Choices: (A) 45 (B) 90 (C) 135 (D) 180
Ground truth: (C) 135

DAPO:
<think>
Given that ABCD is a rectangle, all angles in a rectangle are 90 degrees.
Therefore, angle A is 90 degrees.
<think>

$\boxed{B}$

VEPO:
<think>
To find the measure of angle $\angle A$ in quadrilateral $ABCD$, we need to understand the properties of the angles in a quadrilateral. The sum of the interior angles of any quadrilateral is always 360°.
Given:
- $\angle ABC = 90^\circ$ (since $\angle B$ is a right angle)
- $\angle BCD = 90^\circ$ (since $\angle C$ is a right angle)

Let's denote the measure of $\angle D$ as x° and the measure of $\angle A$ as $3x^\circ$.
The sum of the angles in quadrilateral $ABCD$ is:

$$\angle A + \angle B + \angle C + \angle D = 360^\circ$$

Substituting the known values:

$$3x + 90 + 90 + x = 360$$

Combining like terms:

$$4x + 180 = 360$$

Subtract 180 from both sides:

$$4x = 180$$

Divide by 4:

$$x = 45$$

Now, we can find $\angle A$:

$$\angle A = 3x = 3\times 45 = 135^\circ$$

Therefore, the measure of $\angle A$ is 135°.
<think>
The correct option is $\boxed{C}$

Figure 6: Case study 2 comparing DAPO and VEPO.