

Entropy Is Not Enough: Unlocking Effective Reinforcement Learning for Visual Reasoning via Vision-Anchored Token Selection

Senjie Jin^{1*}, Peixin Wang^{1*}, Boyang Liu^{1*}, Xiaoran Fan^{1*},
Shuo Li¹, Zhiheng Xi¹, Jiazheng Zhang¹, Yuhao Zhou¹,
Tao Gui^{1†}, Qi Zhang¹, Xuanjing Huang¹

¹College of Computer Science and Artificial Intelligence, Fudan University
{sjjin24, peixinwang25, boyangliu25}@m.fudan.edu.cn, tgui@fudan.edu.cn

Abstract

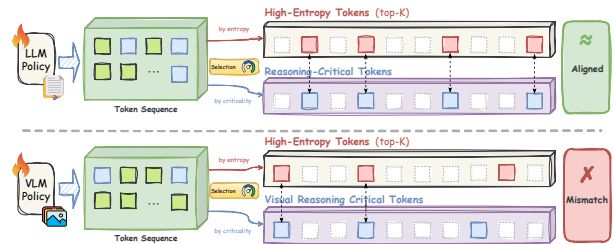
While token-level entropy is commonly recognized as effective for credit assignment in text-only reinforcement learning with verifiable rewards (RLVR), it remains unclear whether this mechanism still holds in visual reasoning. Our controlled study shows that this mechanism collapses in visual reasoning due to the omission of vision-sensitive tokens with naturally low entropy. Although existing multi-modal RL methods increasingly acknowledge the importance of visual perception, they struggle to satisfy the inherent demand for interleaving precise perceptual grounding with semantic reasoning, either lacking systematic visual measurements or overlooking that token entropy primarily drives semantic exploration. To address this, we introduce **VEPO** (Vision-Entropy token-selection for Policy Optimization), an effective RL framework explicitly integrating *visual sensitivity* with *token entropy* via an information-theoretically motivated multiplicative coupling, where VEPO redirects gradient credit toward tokens which are simultaneously visually grounded and highly informative. Extensive experiments demonstrate VEPO’s leading performance, consistently outperforming the strongest RL baseline at 7B-scale and 3B-scale. Ablations further substantiate the soundness of our method¹.

1 Introduction

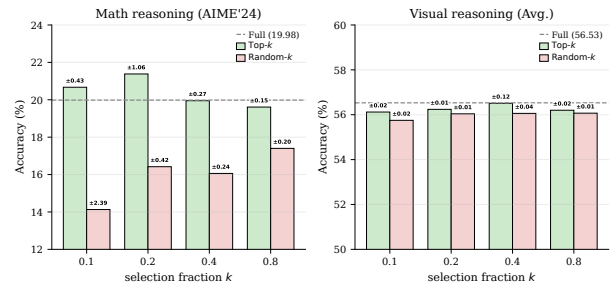
Reinforcement learning with verifiable rewards (RLVR) (Lambert et al., 2024; Shao et al., 2024; Yu et al., 2026) has rapidly emerged as a pivotal paradigm for incentivizing complex reasoning abilities in Large Language Models (LLMs). A fundamental challenge is *how to effectively allocate credit across individual tokens within long reasoning trajectories* (Zhong et al., 2024; Wang et al., 2026). Recent work in text-only RL converges on a

* Equal contribution. † Corresponding authors.

¹<https://github.com/Leonnnnnn929/VEPO>



(a) Entropy selection succeeds in LLMs but collapses in VLMs, omitting vision-critical tokens with low entropy.



(b) Performance of different token-selection mechanisms on math/visual reasoning tasks.

Figure 1: The entropy-based token selection mechanism collapses on visual reasoning: A small subset of high-entropy token training outperforms Full and substantially exceeds Random-k on math reasoning. But on visual reasoning, entropy-based selection only marginally beats Random-k and falls below Full.

compelling empirical regularity: a small fraction of high-entropy tokens, dubbed *forking tokens* (Wang et al., 2026; Bigelow et al., 2025), act as pivotal decision points and account for nearly all of the performance gains in RLVR (Wang et al., 2026); follow-up analyses corroborate that sustained gains concentrate on updates at these high-entropy positions (He et al., 2026). Token-level entropy has therefore been used as an uncertainty signal to modulate credit assignment (Wang et al., 2025c). At the policy level, entropy is highly predictive of downstream performance (Cui et al., 2025), and its collapse is accompanied by a loss of exploration (Cui et al., 2025; Xi et al., 2026).

While this entropy-driven mechanism proves

highly effective for text-only reasoning tasks, whether it generalizes to visual reasoning remains a critical open question (Li et al., 2025). We investigate this through a controlled study (Section 2.1) and observe a striking divergence: the entropy mechanism collapses once the visual modality is introduced (Figure 1b). On math reasoning (AIME’24), we reproduce the prior findings (Wang et al., 2026), yet the identical mechanism underperforms the full-token GRPO baseline on visual reasoning tasks, even yielding marginal gains over the random selection strategy. We further probe the underlying cause and identify a previously uncharacterized phenomenon (Section 2.2): the pure entropy-based token selection inadequately captures genuinely vision-sensitive tokens (Figure 2) that possess naturally low entropy (Favero et al., 2024; Leng et al., 2024). Quantitatively, at $k=0.2$, top-entropy selection recovers merely 59% of tokens flagged by the vision-sensitive metrics, thereby we argue that token entropy alone remains insufficient for identifying the critical tokens in visual reasoning. Qualitative cases are deferred to Figure 2(c)(d) and Appendix J.

Unlike text-only reasoning tasks, visual reasoning intrinsically requires interleaving precise perceptual grounding with semantic reasoning (Gu et al., 2025; Su et al., 2025b). While recent work in multimodal RL (Liu et al., 2025b; Wang et al., 2025d; Li et al., 2026b) has increasingly recognized the importance of visual perception, they either rely on a single indirect visual proxy (e.g., hidden-state similarity) rather than a distributional measurement of visual dependency (Li et al., 2026b), or do not couple visual signals with the fact that entropy primarily serves as a proxy of semantic exploration (Huang et al., 2025; Ye et al., 2026), which visual modalities cannot effectively resolve.

Motivated by above analysis, we propose **VEPO** (Vision-Entropy token-selection for Policy Optimization), an effective token-level credit-assignment RL framework that explicitly couples *vision-sensitivity* with *token entropy*, redirecting gradient credit toward tokens that serve a true functional purpose in visual reasoning. Concretely, VEPO performs a counterfactual forward pass with a noise-perturbed image and quantifies token-level visual dependency by multiplicatively coupling two complementary signals: (i) the Jensen–Shannon divergence (JSD) (Menéndez et al., 1997) for distributional shift, and (ii) the absolute entropy gap $|\Delta H_t|$ for direction-agnostic shifts. This joint formulation

is conceptually inspired by the *aleatoric–epistemic* uncertainty decomposition (Houlsby et al., 2011; Depeweg et al., 2018; Kendall and Gal, 2017), establishing distributional divergence and predictive confidence as complementary axes. Furthermore, the multiplicative coupling acts as a noisy-OR-inspired (Xu et al., 2026; Pearl, 2014), guaranteeing the comprehensive capture of visually sensitive signals by preventing the omission of tokens that spike along either axis (Appendix G). Ultimately, by modulating this fused signal with token entropy, VEPO ensures policy updates are driven exclusively by *visual forking tokens* that are simultaneously visually grounded and highly informative.

We evaluate VEPO with Qwen2.5-VL-7B/3B-Instruct (Bai et al., 2025). Experimental results demonstrate that VEPO achieves leading performance at the 7B scale, outperforming the top-entropy baseline by +2.28, with consistent gains extending to the 3B scale. Moreover, we conduct comprehensive ablation studies and analysis (Section 5) to validate the soundness and effectiveness of our proposed method.

Our contributions are summarized as follows:

- We provide a comprehensive diagnosis revealing why the top-entropy token selection mechanism collapses in visual reasoning, attributing to its **inherent omission of visually grounded tokens**.
- We propose **VEPO**, a novel token-level credit-assignment framework that integrates visual sensitivity and token entropy to achieve a **visually grounded and highly informative** optimization.
- Extensive experiments across seven benchmarks and two model scales, alongside comprehensive ablation studies, demonstrate the effectiveness and soundness of VEPO.

2 Preliminary Experiments

Recent works on math reasoning have uncovered a striking empirical observation: emphasizing high-entropy tokens yields non-trivial benefits (Wang et al., 2026; He et al., 2026; Wang et al., 2025b). Building upon this, we reproduce the experiments and further investigate whether this mechanism generalizes to visual reasoning tasks. Our empirical findings and analyses are elaborated in the subsequent sections, while detailed experimental setup is deferred to Appendix A.

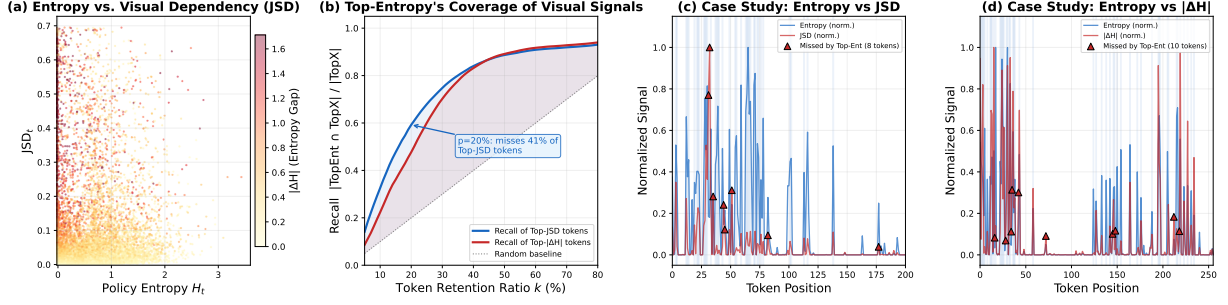


Figure 2: (a) Many high-JSD / high- $|\Delta H|$ tokens lie in the low-entropy region. (b) Top-Entropy selection misses 41% of Top-JSD tokens at $k = 20\%$, alongside a comparable proportion of Top- $|\Delta H|$ tokens. (c,d) Tokens flagged by JSD / $|\Delta H|$ (red triangles) frequently fall in entropy valleys and are missed by Top-Entropy selection.

2.1 The entropy mechanism collapses on visual reasoning

The entropy-based token-selection mechanism collapses when transferred from text to visual reasoning (Figure 1). On math reasoning (AIME’24), selecting a small fraction of high-entropy tokens (Top-10%/20%) exceeds the Full-token baseline, while substantially outperforming Random- k selection across all fractions. This empirically verifies the conjecture presented in (Wang et al., 2026). In contrast, on visual reasoning, we find: 1) Top- k entropy selection consistently falls below the Full-token baseline and yields only marginal gains over Random- k selection. 2) Both the absolute performance and the Top/Random selection gaps remain consistent regardless of the fractions k . These results indicate that token entropy fails to precisely identify critical reasoning decision points once visual evidence is introduced. Detailed results are deferred to Appendix B.

2.2 Unpacking the collapse: The deficiency of visual signals

Drawing inspiration from recent studies that highlight visual evidence during reasoning (Jian et al., 2025; Huang et al., 2025), we hypothesize that high-entropy tokens are sometimes misaligned with critical, visually-grounded decision points. To probe this, we measure two token-level visual dependence signals: the Jensen–Shannon divergence (JSD) (Menéndez et al., 1997) and the entropy gap $|\Delta H_t|$ (Detailed definitions are provided in Section 3.1) between the model’s predictive distribution with and without the image noise.

We make several intriguing observations (illustrated in Figure 2): First, a substantial fraction of high-JSD and high- $|\Delta H|$ tokens concentrates in the low-entropy region (Figure 2(a)). Mechanistically, visual evidence rules out distractor con-

tinuations and sharpens the posterior, so visually-grounded tokens sometimes become more confident, manifesting as lowered entropy (Favero et al., 2024; Leng et al., 2024). Meanwhile, high-entropy tokens are driven more by linguistic indeterminacy (Bigelow et al., 2025; Park and Lei, 2026). Second, Figure 2(b) quantifies the representational gap: at retention ratio $k = 0.2$, Top-Entropy selection recovers a mere 59% of Top-JSD tokens, and similarly, a comparable proportion of Top- $|\Delta H|$ tokens. As k grows, this recall scales in lockstep with the accuracy curve shown in Figure 1(b), gradually approaching the full-token baseline. Performance peaks at $k = 0.4$, where coverage is sufficient and noise remains bounded. At $k = 0.8$, the marginal gain vanishes as the selection is diluted with linguistically-uncertain but visually-irrelevant tokens. Figure 2(c)(d) are two qualitative cases demonstrate how purely entropy-based selection inadequately captures vision-sensitive tokens.

3 Method: VEPO

Motivation. From our findings and analysis in Section 2, we argue that token entropy alone is insufficient to identify critical *forking tokens* in visual reasoning. Following recent works that emphasize visual evidence in the reasoning process (Liu et al., 2025b; Yang et al., 2026; Ou et al., 2026; Li et al., 2026a), we aim to highlight the role of visual signals in evaluating token uncertainty. Hence, we propose **VEPO**, an effective RL training algorithm driven by a joint visual sensitivity and entropy token-selection mechanism. We elaborate on the metrics used to measure token visual sensitivity in Section 3.1, followed by our selection mechanism and training method in Section 3.2 and Section 3.3, all of which are illustrated in Figure 3.

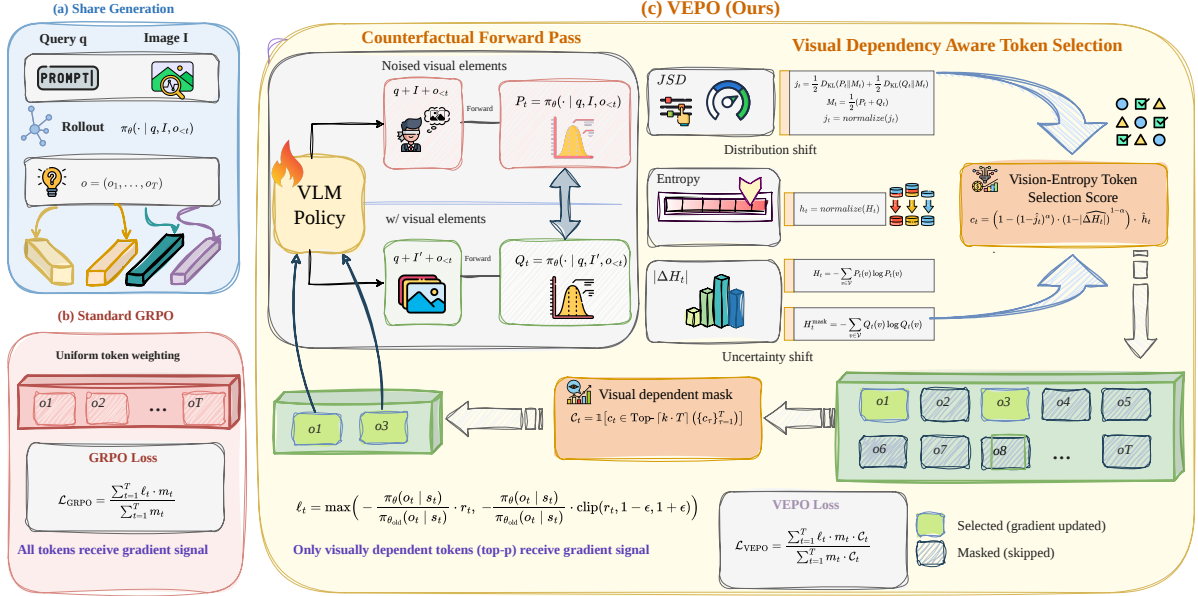


Figure 3: The main training pipeline. VEPO performs a counterfactual forward pass with a noise perturbed image and quantifies token-level visual dependency by multiplicatively coupling two complementary signals: **JSD** and $|\Delta H_t|$, which are then modulated with token **entropy** for simultaneously visually grounded and informative optimization. The detailed algorithm is in Appendix I.

3.1 Token-Level Visual Sensitivity

Given a vision-language policy model parameterized by π_θ , as shown in Figure 3, we perform two forward passes for each generated response $o = (o_1, \dots, o_T)$, yielding a pair of symmetric full-vocabulary next-token distributions at each position t :

$$\begin{aligned} P_t(\cdot) &= \pi_\theta(\cdot \mid q, I, o_{<t}), \\ Q_t(\cdot) &= \pi_\theta(\cdot \mid q, I', o_{<t}), \end{aligned} \quad (1)$$

where P_t and Q_t denote the token distributions conditioned on the original image I and the noisy image I' , respectively, with the same response prefix $o_{<t}$. Their discrepancy therefore isolates the effect of visual corruption. To quantify this distributional change, we use the Jensen–Shannon divergence (JSD) (Menéndez et al., 1997):

$$\text{JSD}_t = \frac{1}{2} D_{\text{KL}}(P_t \parallel M_t) + \frac{1}{2} D_{\text{KL}}(Q_t \parallel M_t), \quad (2)$$

where

$$M_t = \frac{1}{2}(P_t + Q_t). \quad (3)$$

Detailed comparisons and analysis of JSD and KL-based variants are provided in section 5.2.

We further compute the entropy gap of the two distributions to characterize the model’s token pre-

dictive uncertainty:

$$H_t = - \sum_{v \in \mathcal{V}} P_t(v) \log P_t(v), \quad (4)$$

$$H_t^{\text{mask}} = - \sum_{v \in \mathcal{V}} Q_t(v) \log Q_t(v), \quad (5)$$

where H_t and H_t^{mask} denote the entropy under the original image and the corrupted image. We define the entropy gap as

$$\Delta H_t = H_t^{\text{mask}} - H_t, \quad (6)$$

whose magnitude $|\Delta H_t|$ measures how strongly visual corruption changes predictive uncertainty at position t . We therefore use $|\Delta H_t|$ as a direction-agnostic signal of visual sensitivity. Further analysis is provided in section 5.2.

3.2 Joint Vision-Entropy for Token Selection

We first define a joint vision–entropy score for token-level credit assignment to select both visually dependent and highly informative tokens. Following prior token-importance formulations, we combine two visual signals with a probabilistic OR (Xu et al., 2026; Pearl, 2014).

Definition 3.1 (Vision-Entropy Token Selection Score). Let $\hat{j}_t \in [0, 1]$, $|\widehat{\Delta H}_t| \in [0, 1]$, and $\hat{h}_t \in [0, 1]$ denote the per-response min-max normalized JSD, absolute entropy gap, and entropy at position

t . We define the score of token t as

$$c_t := \underbrace{\left(1 - (1 - \hat{j}_t)^\alpha (1 - |\widehat{\Delta H}_t|)^{1-\alpha}\right)}_{g_t} \cdot \hat{h}_t, \quad (7)$$

where $\alpha \in [0, 1]$ controls the trade-off between JSD and $|\Delta H_t|$.

Here g_t is a joint visual-dependency score that combines two complementary signals, $\hat{j}_t$ and $|\widehat{\Delta H}_t|$. By eq. (7), it increases with either signal and peaks when both are strong. The entropy term $\hat{h}_t$ then acts as a modulation factor, constrained by the visual signal, ensuring that we prioritize tokens that are both visually sensitive and highly informative. We provide an information-theoretic interpretation of this formulation in Appendix G. Detailed ablations are provided in Section 5.1.

As shown in Figure 3, we then select the top- k fraction of tokens within each response. For a response of length T , we define the binary mask:

$$\mathcal{C}_t = 1[c_t \in \text{Top-}[kT]](\{c_\tau\}_{\tau=1}^T), \quad (8)$$

where only tokens with $\mathcal{C}_t = 1$ are used for the policy gradient update.

3.3 Reinforcement Enhanced Visual Reasoning

Standard GRPO (Shao et al., 2024) computes the policy gradient loss over all valid response tokens. We modify it by restricting the update to the selected subset indicated by $\mathcal{C}_t$:

$$\mathcal{L}_{\text{VEPO}} = \frac{\sum_{t=1}^T \ell_t \cdot m_t \cdot \mathcal{C}_t}{\sum_{t=1}^T m_t \cdot \mathcal{C}_t}, \quad (9)$$

where m_t is the standard response mask. The per-token clipped policy gradient loss is:

$$\begin{aligned} \ell_t &= \max(-A \cdot r_t, -A \cdot \hat{r}_t), \\ \hat{r}_t &= \text{clip}(r_t, 1-\epsilon, 1+\epsilon), \end{aligned} \quad (10)$$

with importance ratio:

$$r_t = \frac{\pi_\theta(o_t | s_t)}{\pi_{\theta_{\text{old}}}(o_t | s_t)}. \quad (11)$$

By replacing m_t with $m_t \cdot \mathcal{C}_t$, VEPO concentrates gradient updates on *visual forking* tokens.

4 Experiment

4.1 Experiment Setup

Dataset and Evaluation. We construct the training set from subsets of Geometry3K (Lu et al.,

2021) and MMK12 (Meng et al., 2025), resulting in approximately 4.2K training samples. We evaluate the model on both in-domain and out-of-domain benchmarks: the Geometry3K and MMK12 validation splits serve as in-domain evaluation, while a set of public visual reasoning benchmarks is used to assess out-of-domain generalization. Detailed benchmark descriptions are provided in Appendix C.

Implementation. We use the EasyR1 framework (Zheng et al., 2025) as the reinforcement learning framework for vision-language model training. The policy models are initialized from Qwen2.5-VL-3/7B-Instruct², followed by fine-tuning on our custom dataset. More implementation details can be found in Appendix C.

Baseline. We compare our method with several baselines: closed-source multimodal models, open-source large-scale VLMs, visual-focused RL fine-tuning methods, and top-entropy token selection methods. For RL-based and entropy-based selection baselines, we retrain these methods under the same 4.2K mixed training set for fair comparison. More details can be found in Appendix C.

4.2 Main Results

The main results are presented in Table 1. We primarily analyze the model performance compared with top-entropy selection methods and effective visual-focused RL fine-tuning methods. We also included the scaling experiments on the 3B model. To assess training stability, we further conduct multi-seed experiments for VEPO at both the 7B and 3B scales, with detailed results reported in Tables 9 and 10.

Comparison with Top-Entropy selection. We first compare VEPO against entropy-based methods. As shown at the bottom of Table 1, both entropy-only variants (top-20% / top-40%) underperform the full-token GRPO baseline on average (56.54 / 56.50 vs. 57.11), corroborating our preliminary observation in Section 2 that token entropy alone is an insufficient selection signal for visual reasoning tasks. By further incorporating the visual-sensitivity criterion, VEPO reaches a **+2.28** gain over the strongest entropy-only variant and outperforms full-token GRPO by **+1.71**. The benefits are more notable on in-domain benchmarks (Geo3K **+4.16**, MMK12 **+4.38**) and transfer consistently to all out-of-domain tasks except for a

²<https://huggingface.co/collections/Qwen/qwen25-vl>

Model / Dataset	Geo3K	MMK12	HalluBench	MathVista	We-Math	MathVerse	MathVision	Avg.
<i>Closed-source Models</i>								
GPT-4o (OpenAI et al., 2024)	37.60	49.40	72.10	62.60	69.50	47.70	33.90	53.26
o1 (OpenAI et al., 2026)	63.20	65.90	74.50	73.90	98.70	57.00	60.30	70.50
Claude-3.7-Sonnet (Anthropic, 2025a)	–	55.30	–	66.80	72.60	52.00	41.30	–
Claude-4-Sonnet (Anthropic, 2025b)	50.40	71.30	74.10	74.70	81.50	51.30	51.10	64.91
Gemini-2.0-Flash (Gemini Team, Google, 2025)	–	65.20	–	70.40	71.40	59.30	41.30	–
Gemini-2.5-Flash (Gemini Team, Google, 2025)	53.70	57.10	72.70	71.10	73.40	49.80	44.90	60.39
<i>Open-source Models</i>								
Qwen2.5-VL-32B (Bai et al., 2025)	50.10	66.80	71.80	74.70	69.10	49.90	40.10	60.36
InternVL2.5-38B (Chen et al., 2025b)	44.30	58.00	56.80	71.90	67.50	49.40	31.80	54.24
Qwen2.5-VL-72B (Bai et al., 2025)	51.20	70.50	72.90	74.80	72.40	57.60	38.10	62.50
QvQ-72B-Preview (Team, 2024)	34.60	61.50	66.40	71.40	65.40	48.20	35.90	54.77
InternVL2.5-78B (Chen et al., 2025b)	47.30	61.60	67.40	72.30	66.30	51.70	32.20	56.97
<i>Visual-focused RL Fine-tuning Methods (7B)</i>								
NoisyRollout-7B (Liu et al., 2025b)	48.42	66.05	71.29	71.70	69.20	49.16	28.91	57.82
PAPo-DAPO-7B (Wang et al., 2025d)	50.08	66.05	72.24	73.10	68.69	48.07	26.23	57.78
VPPO-7B (Huang et al., 2025)	50.42	69.02	70.14	72.10	69.43	48.21	28.19	58.22
R1-ShareVL (Yao et al., 2026)	46.59	65.68	69.93	68.10	70.17	47.39	27.27	56.45
<i>Our Method</i>								
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	35.94	53.66	63.51	66.87	62.13	43.27	25.10	50.07
+ GRPO	49.42	65.18	70.45	71.60	67.18	48.53	27.40	57.11
+ GRPO (Top 20% high entropy)	47.59	64.93	68.56	70.40	67.13	48.25	28.95	56.54
+ GRPO (Top 40% high entropy)	46.76	63.44	69.09	71.40	68.45	48.86	27.53	56.50
+ VEPO*	51.75 (± 0.76)	69.31 (± 0.47)	71.12 (± 0.63)	72.37 (± 0.32)	69.62 (± 0.09)	49.15 (± 1.18)	28.42 (± 0.20)	58.82 (± 0.09)
Δ vs. GRPO (Top 20% high entropy)	(+4.16)	(+4.38)	(+2.56)	(+1.97)	(+2.49)	(+0.90)	(−0.53)	(+2.28)

Table 1: The main experimental results across multiple visual math and hallucination benchmarks. Models are trained on the Geo3K and MMK12 datasets, which serve as in-domain evaluations, while the remaining datasets assess out-of-domain generalization. Our proposed VEPO (bottom) demonstrates superior overall performance compared to the corresponding baselines. Best and second-best results are determined only among the visual-focused RL fine-tuning methods and our methods, and are highlighted in **bold** and underlined, respectively. Results for models marked with * are reported as mean ($\pm$ standard deviation) over multiple training runs with different random seeds.

marginal -0.53 drop on MathVision. These results suggest that visual sensitivity serves as an effective complementary signal to token entropy, helping to identify the effective decision points in visual reasoning.

Comparison with Visual-focused RL methods. We further compare VEPO with representative visual-focused RL fine-tuning methods. As shown in the middle of Table 1, VEPO attains the highest average of **58.82**, consistently surpassing all baselines. Among these, VPPO (Huang et al., 2025) is the most competitive one, leveraging the KL divergence between clear and perturbed images for token selection. While we acknowledge the crucial role of visual signals in visual reasoning, the reasoning trajectory itself remains predominantly textual. Broad exploration (Cheng et al., 2026) along this textual reasoning chain is essential for pushing beyond performance ceiling, token entropy therefore remains an indispensable signal. This motivation lies at the heart of VEPO, which couples visual sensitivity with token entropy through a multiplicative scoring mechanism. A more comprehensive analysis is deferred to Section 5.1.

Results across Different Model Scales. To verify VEPO’s scalability, we evaluate it on Qwen2.5-

VL-3B-Instruct (Table 6), yielding consistent conclusions. Details are deferred to Appendix D. Furthermore, VEPO-7B outperforms several open-source multimodal models that are approximately $5\text{--}11\times$ larger in terms of average performance, including InternVL2.5-38B, InternVL2.5-78B, and QvQ-72B-Preview. On MMK12, VEPO (69.31) surpasses Qwen2.5-VL-32B (66.80) and both InternVL2.5 variants (58.00/61.60), while remaining competitive with Qwen2.5-VL-72B (70.50). On MathVista, VEPO (72.37) surpasses InternVL2.5-38B (71.90), InternVL2.5-78B (72.30), and QvQ-72B-Preview (71.40). These results further corroborate the effectiveness of VEPO.

5 Ablation Study and Analysis

5.1 Ablation Study

Signals of VEPO Token Selection Mechanism.

We first ablate the three signals in our joint selection mechanism (Equation (7)). As shown in Table 2, removing any signal degrades average performance, confirming their complementary roles. The two visual signals (JSD and $|\Delta H_t|$) contribute the most, with their removal causing nearly identical drops of -2.68 and -2.71 , respectively. Although omitting the entropy signal $\hat{h}_t$ yields a milder av-

Variant	Geo3K	MMK12	HalluBench	MathVista	We-Math	MathVerse	MathVision	Avg.
Full VEPO*	51.75	69.31	71.12	72.37	69.62	49.15	28.42	58.82
<i>Ablation on VEPO token selection signals</i>								
w/o JSD	46.26	66.67	69.40	68.70	66.84	47.82	27.30	56.14
w/o Entropy Gap	46.42	65.43	68.35	70.80	66.67	48.32	26.78	56.11
w/o Entropy	47.75	69.02	70.35	72.30	69.48	48.98	27.34	57.89
<i>Ablation and Analysis on variants of VEPO token selection mechanism</i>								
Replacing JSD with KL	50.42	69.27	69.30	71.60	68.22	49.87	27.73	58.06
Bottom 80%	41.60	60.97	68.45	66.40	67.53	43.89	26.38	53.60
$\max(\Delta H_t, 0)$	50.42	67.78	69.71	71.40	68.67	49.74	28.42	58.02
Additive Fusion	49.75	70.01	68.66	71.60	67.99	50.23	28.68	58.13

Table 2: Ablation and analysis of VEPO token selection variants. **Full VEPO** denotes the complete joint selection and is reported as the multi-seed average, consistent with the main results. The upper block ablates each signal individually (JSD, entropy gap, or entropy), while the lower block examines alternative formulations of the VEPO’s selection mechanism. **Additive Fusion** replaces the multiplicative aggregation of two visual signals with linear interpolation: $g_t = \alpha \hat{j}_t + (1 - \alpha) |\Delta H_t|$. Results marked with * are averaged over multiple training runs with different random seeds.

erage decline (-0.93), we find it severely impacts benchmarks dominated by long textual reasoning chains (Geo3K, MathVision) while remaining neutral on perception-heavy ones (MathVerse, MathVista). This empirically validates that entropy and visual sensitivity are *complementary signals* rather than competing alternatives.

Visual Sensitivity α and Token Fraction k . We further investigate the effects of two key hyperparameters: the visual sensitivity balancing coefficient α and the token retention ratio k . As illustrated in Figure 4, we summarize several key findings: 1) The optimal configuration is ($k=0.2, \alpha=0.7$), yielding the highest average score. 2) The dense retention ratio $k=0.4$ uniformly underperforms the sparser settings across all α , confirming that retaining too many tokens introduces weakly grounded positions into the policy update, consistent with (Wang et al., 2026). 3) Performance peaks at $\alpha=0.7$ and degrades sharply at either extreme ($\alpha=0.1$ or $\alpha=0.9$), empirically confirming that JSD and $|\Delta H_t|$ serve as *complementary visual signals*, with JSD playing a more prominent role. 4) Interestingly, the optimal k adapts to the strength of α . When α is small to moderate ($0.1 \sim 0.5$), $k=0.3$ outperforms $k=0.2$, as a weaker JSD signal benefits from a broader token pool. Conversely, at $\alpha \geq 0.7$, $k=0.2$ becomes optimal as the stronger JSD signal enables reliable selection from a sparser set. Detailed results on each benchmark are deferred to Appendix E.1.

5.2 Analysis

Comparison with visual signal variants. We compare our visual signal formulation against three

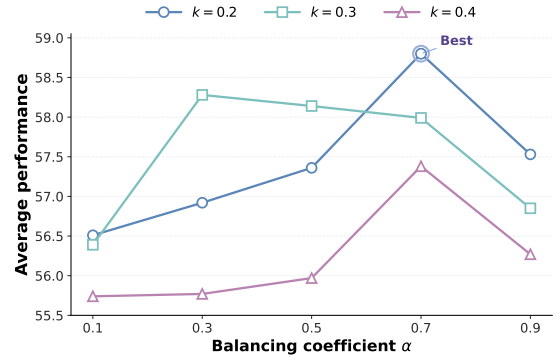


Figure 4: Ablations on the balancing coefficient α and the token selection ratio k .

information-theoretically motivated alternatives. (1) Replacing JSD with KL, the metric utilized in VPPO (Huang et al., 2025), decreases the average by 0.76 points. Unlike VPPO’s trajectory-level advantage reweighting, VEPO employs token-level selection driven by a multiplicative coupling $v_t = g_t \cdot \hat{h}_t$. This makes unbounded signals like KL highly susceptible to outlier tokens, whereas JSD remains bounded and symmetric (Menéndez et al., 1997; Lin, 1991). (2) Substituting $|\Delta H_t|$ with $\max(\Delta H_t, 0)$ incurs a 0.80 drop. This confirms that tokens whose entropy decreases upon visual conditioning (where the image successfully disambiguates an uncertain prediction) serve as equally informative cues for visual dependency (Nam et al., 2025). (3) Replacing the multiplicative aggregation with additive fusion drops by 0.69 points. The multiplicative form implements a noisy-OR-inspired weighted aggregation (Pearl, 2014) that saturates when either signal is strong, while additive fusion under-learns non-additive interactions (Jayakumar et al., 2020). These results validate VEPO’s bounded, direction-agnostic, and multi-

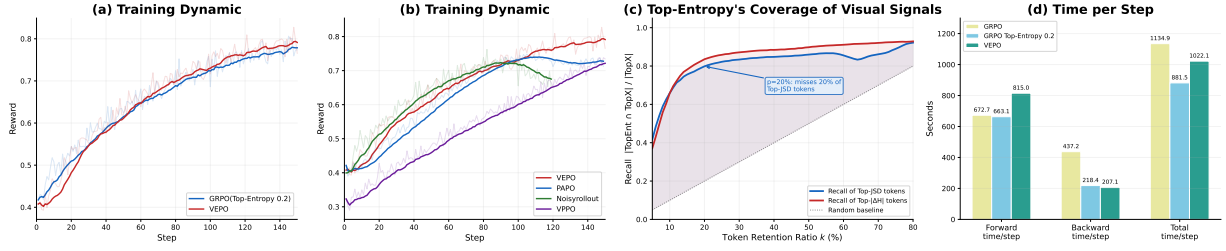


Figure 5: (a)(b) The training dynamics of VEPO and top-entropy and visual-focused RL methods (VPPO (Huang et al., 2025) removes format reward, we follow the default setting). (c) At $k=20\%$, VEPO covers more vision-sensitivity tokens, reducing miss rate to 20%. (d) Per-step wall-clock time of VEPO and GRPO baselines.

plicatively aggregated visual signals as information-theoretically motivated selection choices. More analysis about the visual perturbation is deferred to Appendix E.2.

Impact of High-JSD, Low-Entropy Tokens. Our diagnostic experiments in Section 2 find that entropy alone is insufficient in visual reasoning. To evaluate whether VEPO-exclusive tokens contribute directly to downstream performance, we define $D = S_{\text{VEPO}} \setminus S_{\text{Entropy}}$, where both methods select 20% of tokens, we remove D from S_{VEPO} and replace it with an equal number of randomly selected tokens from the unselected 80%, preserving the total update budget. As reported in Table 11, this subset constitutes approximately 14.57% of the tokens selected by VEPO. Replacing these tokens reduces the average performance from 58.82 to 57.30. This result demonstrates that the additional positions identified by VEPO are functionally important rather than incidental selection differences.

Training dynamic and efficiency. Figure 5 (a) and (b) illustrate the training dynamics of VEPO and corresponding baselines. We observe that visual-focused RL is prone to premature convergence and optimization instability, whereas our method maintains stable training dynamics, yields higher rewards, thereby exhibits superior scalability (Khatri et al., 2025). Figure 5(c) explains this gap mechanistically: at $k=0.2$, top-entropy selection misses $\sim 40\%$ of vision-sensitive (Top-JSD / $|\Delta H_t|$) tokens. In contrast, VEPO reduces this to just 20%. By incorporating more visually sensitive tokens, VEPO enables the model to strike an optimal balance between perceptual grounding and reasoning informativeness. Figure 5 (d) reports per-step wall-clock time: by restricting gradient updates to top- k tokens, VEPO is $\sim 10\%$ faster than Full-token GRPO and $\sim 16\%$ slower than Top-Entropy selection, which is a modest trade-off between computation and performance.

Qualitative Analysis. Figure 6 and Figure 7 in Appendix J illustrate the distinct token selection patterns between VEPO and the top-entropy mechanism. Both methods identify canonical reasoning anchors (“Initialize”, “condition”), confirming a shared foundational baseline. Beyond this base, entropy-only selections drift to LaTeX scaffolding (“\frac”, “\boxed”) and discourse fillers (“follows”, “we”), where uncertainty stems from linguistic variability rather than visual dependency. In contrast, VEPO uniquely recovers visually-grounded content tokens, like variables and operators tied to the code (“a”, “+”) and condition keywords referencing the flowchart (“greater”, “when”). This demonstrates VEPO’s ability to accurately redirect credit toward tokens that genuinely ground the textual reasoning in the visual input.

6 Related Work

Entropy Mechanism and Credit Assignment .

In text-only reinforcement learning (RL), entropy is a well-formalized proxy for policy performance (Cui et al., 2025), where a small minority of high-entropy tokens dominates exploration and provides the primary gradient signals (Wang et al., 2026; Yu et al., 2026; Xi et al., 2026). Whether this principle transfers to the multimodal setting, however, remains underexplored (Li et al., 2025): in visual reasoning, token uncertainty arises from heterogeneous sources spanning reasoning (Chung et al., 2025), visual perception and image noise (Fang et al., 2026). Existing multimodal RL frameworks either conflate general high entropy with genuine visual dependency (Ou et al., 2026; Chen et al., 2026), or, as in VPPO (Huang et al., 2025) and PGPO (Ye et al., 2026), quantify per-token visual sensitivity via KL divergence between distributions conditioned on the original and a perturbed image (Huang et al., 2025), whereas PAPO (Wang et al., 2025d) introduces visual KL as an auxiliary perception objective. We bridge this gap with VEPO,

which couples the vision-sensitivity with entropy to identify pivotal token-level decision points in visual reasoning.

Visual Perception Enhancement. Unlike the strong reasoning shown in text-only domains such as math and code (Guo et al., 2025; OpenAI et al., 2026; Chen et al., 2025a), MLLMs still lag on complex visual reasoning, a gap recently attributed to inadequate visual perception and grounding rather than reasoning itself (Ou et al., 2026). Current multimodal RL approaches fall into three lines: (i) shaping visual attention via auxiliary rewards (Jian et al., 2025; Ou et al., 2026), auxiliary training objectives (Wang et al., 2025d; Li et al., 2026a), or auxiliary visual-contrast signals (Wang et al., 2024a; Liu et al., 2025a; Huang et al., 2025; Ye et al., 2026; Li et al., 2026b); (ii) multi-stage workflows or curated visually-grounded data (Yang et al., 2026; Wang et al., 2025a); and (iii) tool-augmented reasoning with external visual operations such as zooming or cropping (Zheng et al., 2026; Hong et al., 2026; Su et al., 2025a). These methods, however, often optimize visual perception separately from semantic exploration or rely on external tool scaffolding. (Liu et al., 2025c) show that correctly localized visual evidence does not necessarily yield correct answers, highlighting the need for effective token-level credit assignment. Motivated by this, VEPO reframes visual grounding as a token-level credit-assignment problem. Rather than relying on a single visual proxy, it jointly captures distributional visual dependency through JSD, predictive-confidence changes through the entropy gap, and reasoning informativeness through token entropy. Its key distinction therefore lies in joint signal composition and credit allocation for visual reasoning.

7 Conclusion

In this paper, we revisit entropy-driven token selection in visual reasoning and show a counterintuitive failure mode: this mechanism sometimes overlooks vision-sensitive tokens that naturally exhibit low entropy but carry crucial visual evidence. Building on this diagnosis, we propose **VEPO**, a novel RL framework that explicitly couples visual sensitivity with token entropy through a multiplicative aggregation, redirecting gradient credit toward tokens that are simultaneously visually grounded and informative. Extensive experiments and ablations confirm VEPO’s effectiveness. It achieves leading

performance at comparable scales, surpassing the entropy-only baseline by +2.28 and +3.11 points on 7B and 3B models. We hope this work motivates further investigation into modality-aware credit assignment in multimodal reasoning.

Limitations

Our work has a few limitations that warrant further investigation. First, the auxiliary forward pass required for visual perturbation introduces additional training overhead, though this could potentially be mitigated by advanced KV-cache reuse. Second, while Gaussian noise is an effective default counterfactual and we explicitly evaluated the no-image condition (Appendix E.2) as an alternative perturbation, current computational limitations restricted our exploration of broader noise variants and sophisticated scheduling techniques like annealing (Liu et al., 2025b). Lastly, while VEPO demonstrates strong performance and we offer an intuitive interpretation for its empirical effectiveness, providing a rigorous theoretical foundation for why visual-sensitive token selection improves multimodal RL optimization is left for future work.

Ethical Considerations

This work uses publicly available datasets and open-source vision-language models for training, as well as proprietary models for comparison and automatic evaluation. We adhere to the ACL Ethics Policy and applicable licenses and terms of use. We carefully examined the data to mitigate potential ethical and privacy risks. We emphasize the responsible use of LLMs to reduce the risk of harmful or biased outputs. Finally, our reinforcement learning experiments require substantial computational resources, resulting in energy consumption and associated environmental impacts.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 62576106, 62521004, 62476061, 62376061, 62506079).

References

Anthropic. 2025a. Claude 3.7 sonnet system card. Technical report, Anthropic.

- Anthropic. 2025b. Claude opus 4 & claude sonnet 4 system card. Technical report, Anthropic.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. 2025. Forking paths in neural text generation. In *International Conference on Learning Representations*.
- Jiangjie Chen, Wenxiang Chen, Jiacheng Du, Jinyi Hu, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Wenlei Shi, Zhihong Wang, Mingxuan Wang, Chenrui Wei, Shufa Wei, Huajian Xin, Fan Yang, Weihao Gao, Zheng Yuan, Tianyang Zhan, and 3 others. 2025a. [Seed-prover 1.5: Mastering undergraduate-level theorem proving via learning from experience](#). *CoRR*, abs/2512.17260.
- Shuang Chen, Hangyu Guo, Yimeng Ye, Shijue Huang, Wenbo Hu, Jiayu Chen, Manyuan Zhang, Haoxi Li, Song Guo, and Nanyun Peng. 2026. [ARES: Multi-modal adaptive reasoning via difficulty-aware token-level entropy shaping](#). In *The Fourteenth International Conference on Learning Representations*.
- 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, and 23 others. 2025b. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *Preprint*, arXiv:2412.05271.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. 2026. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30377–30385.
- Jiwan Chung, Neel Joshi, Pratyusha Sharma, Youngjae Yu, and Vibhav Vineet. 2025. What mllms learn about when they learn about multimodal reasoning: Perception, reasoning, or their integration? *arXiv preprint arXiv:2510.01719*.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.
- Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR.
- Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. 2026. Enhancing vision-language model reliability with uncertainty-guided dropout decoding. *Advances in Neural Information Processing Systems*, 38:149193–149218.
- Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.
- Gemini Team, Google. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *arXiv preprint arXiv:2507.06261*.
- Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. 2025. [Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning](#). *CoRR*, abs/2510.27492.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. [Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14375–14385. IEEE.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shiron Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Yuhang He, Haodong Wu, Siyi Liu, Hongyu Ge, Hange Zhou, Keyi Wu, Zhuo Zheng, Qihong Lin, Zixin Zhong, and Yongqi Zhang. 2026. Rethinking token-level credit assignment in rlvr: A polarity-entropy analysis. *arXiv preprint arXiv:2604.11056*.
- Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and XingYu. 2026. [Deepeyesv2: Toward agentic multimodal model](#). In *The Fourteenth International Conference on Learning Representations*.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.
- Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. 2025. [Spotlight on token perception for multimodal reinforcement learning](#). *CoRR*, abs/2510.09285.

- Siddhant M Jayakumar, Wojciech M Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. 2020. Multiplicative interactions and where to find them. In *International conference on learning representations*.
- Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. 2025. [Look again, think slowly: Enhancing visual reflection in vision-language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 9251–9270. Association for Computational Linguistics.
- Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.
- Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. 2025. [The art of scaling reinforcement learning compute for llms](#). *CoRR*, abs/2510.13786.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.
- 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.
- Bangzheng Li, Jianmo Ni, Chen Qu, Ian Miao, Liu Yang, Xingyu Fu, Muhao Chen, and Derek Zhiyuan Cheng. 2026a. [Reinforced attention learning](#). *CoRR*, abs/2602.04884.
- Shuo Li, Jiajun Sun, Zhihao Zhang, Xiaoran Fan, Senjie Jin, Hui Li, Yuming Yang, Junjie Ye, Lixing Shen, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. [The role of entropy in visual grounding: Analysis and optimization](#). *Preprint*, arXiv:2512.06726.
- Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, Zhaojie Liu, Qibin Hou, and Ming-Ming Cheng. 2026b. Rethinking token-level policy optimization for multimodal chain-of-thought. *arXiv preprint arXiv:2603.22847*.
- Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.
- Rui Liu, Dian Yu, Tong Zheng, Runpeng Dai, Zongxia Li, Wenhao Yu, Zhenwen Liang, Linfeng Song, Haitao Mi, Pratap Tokekar, and 1 others. 2025a. [Vogue: Guiding exploration with visual uncertainty improves multimodal reasoning](#).
- Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025b. [Noisyrollout: Reinforcing visual reasoning with data augmentation](#). *Preprint*, arXiv:2504.13055.
- Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhan Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoît Dumoulin, and Hanghang Tong. 2025c. [Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms](#). *CoRR*, abs/2510.17771.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6774–6786. Association for Computational Linguistics.
- María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhao 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*.
- Heejeong Nam, Jinwoo Ahn, Keummin Ka, Jiwan Chung, and Youngjae Yu. 2025. [Vague: visual contexts clarify ambiguous expressions](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1537–1547.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard

- Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 246 others. 2026. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- Siqu Ou, Tianrui Wan, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. 2026. [Do mllms really see it: Reinforcing visual attention in multimodal llms](#). *CoRR*, abs/2602.08241.
- Seunghyun Park and Yuanyuan Lei. 2026. Where reasoning breaks: Logic-aware path selection by controlling logical connectives in llms reasoning chains. *arXiv preprint arXiv:2604.20564*.
- Judea Pearl. 2014. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.
- Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. [We-math: Does your large multimodal model achieve human-like mathematical reasoning?](#) *Preprint*, arXiv:2407.01284.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. 2025a. [Openthinking: Learning to think with images via visual tool reinforcement learning](#). *Preprint*, arXiv:2505.08617.
- Zhaochen Su, Peng Xiang, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. (May) Fung. 2025b. [Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers](#). *CoRR*, abs/2506.23918.
- Qwen Team. 2024. [Qvq: To see the world with wisdom](#).
- Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdp: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2025a. [V1-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning](#). *Preprint*, arXiv:2504.08837.
- Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, and Guorui Zhou. 2025b. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr. *arXiv preprint arXiv:2507.15778*.
- Jiawei Wang, Jiakai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. 2025c. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265*.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024b. [Measuring multimodal mathematical reasoning with math-vision dataset](#). *Preprint*, arXiv:2402.14804.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2026. [Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. 2025d. [Perception-aware policy optimization for multimodal reasoning](#). *CoRR*, abs/2507.06448.
- Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Xun Deng, Zhihao Zhang, Honglin Guo, Zhikai Lei, Miao Zheng, Guoteng Wang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. 2026. [BAPO: Stabilizing off-policy reinforcement learning for LLMs via balanced policy optimization with adaptive clipping](#). In *The Fourteenth International Conference on Learning Representations*.
- Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. 2026. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*.
- Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. 2026. [Look-back: Implicit visual re-focusing in MLLM reasoning](#). In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 11694–11702. AAAI Press.
- Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and Jiaying Huang. 2026. [R1-shareVL: Incentivizing reasoning capabilities of multimodal large language models via share-GRPO](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, Ziming Li, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. 2026. [Not all tokens see equally: Perception-grounded policy optimization for large vision-language models](#). *CoRR*, abs/2604.01840.

- Qiyang 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.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. [MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems?](#) In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, Lecture Notes in Computer Science, pages 169–186. Springer.
- Yaowei Zheng, Juntong Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. 2025. Easyr1: An efficient, scalable, multi-modality rl training framework. <https://github.com/hiyouga/EasyR1>.
- Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. 2026. [Deepeyes: Incentivizing "thinking with images" via reinforcement learning](#). In *The Fourteenth International Conference on Learning Representations*.
- Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. 2024. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*.

A Preliminary Experiment Setup

A.1 Setup for Preliminary Experiments in Section 2.1

Setup. We compare three token selection strategies under the 80/20 rule (Wang et al., 2026) in the standard GRPO(Shao et al., 2024) setting: (1) **Full-token**, where all valid response tokens participate in the policy gradient update; (2) **High-entropy**, which retains only the top- k fraction of tokens ranked by policy entropy; and (3) **Random**, which keeps each valid token independently with probability p via Bernoulli sampling. For the high-entropy and random strategies, we sweep the retention ratio over $k \in \{0.1, 0.2, 0.4, 0.8\}$, and compare them against the full-token baseline ($k = 1.0$).

We conduct experiments on both LLMs and VLMs. For the LLM setting, we train Qwen2.5-VL-7B on a 54.4K mathematical reasoning dataset using the DAPO (Yu et al., 2026), with 8 H800 GPUs, a batch size of 256, 8 rollouts per prompt, and a maximum response length of 10,240 tokens. We evaluate on AIME’24. For the VLM setting, we follow the NoisyRollout setup for fair comparison and train Qwen2.5-VL-7B-Instruct on MMK12-mini (2.1K samples) using EasyR1(Zheng et al., 2025), with 8 GPUs, a global batch size of 128, 12 rollouts per prompt, and a maximum response length of 2,048 tokens, while keeping the vision encoder frozen throughout training. We evaluate all settings on MathVerse (Zhang et al., 2024), Math-Vision (Wang et al., 2024b), MathVista (Lu et al., 2024), We-Math (Qiao et al., 2024), and HallusionBench (Guan et al., 2024). Both the LLM and VLM settings use a learning rate of 1×10^{-6} and token-mean loss aggregation.

A.2 Setup for Preliminary Experiments in Section 2.2

Experiment Setup. To empirically investigate the relationship between policy entropy and visual dependency at the token level, we conduct an inference-time analysis using the same setup as Appendix A.1. For each sample, we perform autoregressive generation with greedy decoding (temperature=0, top- $k=1.0$) and a maximum generation length of 256 tokens.

Visual Dependence Metrics. During generation, we maintain two parallel decoding streams that share the same generated token sequence but differ in visual input: (1) a *factual* stream conditioned

on the original image I , and (2) a *counterfactual* stream conditioned on a perturbed image I' . At each decoding step, we record three token-level signals: (i) the Jensen-Shannon divergence (JSD) between the factual distribution P_t and the counterfactual distribution Q_t over the full vocabulary, (ii) the Shannon entropy H_t of the factual distribution, and (iii) the entropy gap

$$\Delta H_t = H_t^{\text{mask}} - H_t,$$

which measures how degraded visual evidence changes model uncertainty.

Analysis Protocol. We aim to quantify how well policy entropy serves as a proxy for visual dependency. For each token selection ratio k , we first construct an entropy-based selection set A_{ent} by taking the top- k fraction of tokens ranked by entropy. We then compare it against two visual-dependency sets, denoted by A_{vis} , which are defined as the top- k tokens ranked by JSD and by $|\Delta H|$, respectively. We measure the overlap using recall, defined as

$$\text{Recall} = \frac{|A_{\text{ent}} \cap A_{\text{vis}}|}{|A_{\text{vis}}|},$$

where $|A_{\text{ent}} \cap A_{\text{vis}}|$ denotes the number of tokens selected by both methods, and $|A_{\text{vis}}|$ is the size of the visual-dependency set.

B More Detailed Results on Preliminary Experiments

Table 3 and Table 4 report preliminary comparisons of different token selection strategies across three independent runs in the LLM and VLM settings, respectively, including full-token training, entropy-based Top- $k\%$ selection, and random selection.

We first examine the LLM setting. The results clearly support the 80/20 rule in language-only reasoning: entropy-based Top- $k\%$ selection consistently outperforms Random- $k\%$, and the gains are most pronounced when k is small. In particular, sparse selection performs best, and both Top-10% and Top-20% surpass the full-token baseline. This suggests that, in LLM reasoning, a small subset of high-entropy tokens indeed contributes disproportionately to exploration and optimization.

In contrast, the VLM setting exhibits a markedly different pattern. Here, entropy-based token selection provides only limited and unstable benefits: its performance is generally close to that of random selection and does not consistently outperform the

Dataset / Setting	Full	Top-0.1	Random-0.1	Top-0.2	Random-0.2	Top-0.4	Random-0.4	Top-0.8	Random-0.8
AIME'24 Turn 1	19.95	21.21	13.40	21.63	16.45	19.96	16.75	19.15	17.50
AIME'24 Turn 2	20.62	19.75	16.30	22.49	15.62	20.58	15.80	20.10	17.90
AIME'24 Turn 3	19.38	21.04	12.70	20.01	17.20	19.30	15.63	19.58	16.80
<i>Avg.(Std.)</i>									
Avg.(Std.)	19.98(±0.62)	20.67(±0.80)	14.13(±1.91)	21.38(±1.26)	16.42(±0.79)	19.95(±0.64)	16.06(±0.60)	19.61(±0.48)	17.40(±0.56)

Table 3: Preliminary comparison of different token selection strategies in the LLM setting across three independent training runs. We compare full-token training, entropy-based Top- k % selection, and random token selection under multiple retention ratios, and additionally report the mean and standard deviation across runs. The results clearly exhibit the 80/20 effect in language-only reasoning: entropy-based selection consistently outperforms random selection and achieves the best performance under sparse token retention.

Dataset / Setting	Full	Top-0.1	Random-0.1	Top-0.2	Random-0.2	Top-0.4	Random-0.4	Top-0.8	Random-0.8
<i>Turn 1</i>									
HalluBench	71.60	68.77	70.91	69.30	69.19	69.30	68.98	69.51	69.19
MathVista	71.10	71.90	70.70	71.10	70.40	71.70	71.50	70.20	70.30
We-Math	66.78	64.71	65.40	66.61	65.63	66.90	66.03	66.49	65.57
MathVerse	47.16	46.90	45.96	46.68	47.79	47.66	46.52	47.21	47.23
MathVision	27.20	27.39	26.32	27.20	27.53	27.73	27.91	28.22	27.62
Avg.	56.77	55.93	55.86	56.18	56.11	56.66	56.19	56.33	55.98
<i>Turn 2</i>									
HalluBench	68.24	70.03	70.98	70.14	69.61	69.61	70.35	69.40	69.30
MathVista	69.90	69.50	69.07	70.10	69.80	71.00	70.30	70.70	70.20
We-Math	68.10	65.00	65.92	66.09	65.98	67.82	64.48	65.23	67.30
MathVerse	46.88	47.84	45.63	47.06	47.01	48.30	46.32	46.88	46.29
MathVision	28.66	28.30	27.77	26.84	27.93	27.48	27.51	27.76	27.35
Avg.	56.36	56.13	55.87	56.05	56.07	56.84	55.79	55.99	56.09
<i>Turn 3</i>									
HalluBench	68.66	69.19	69.40	69.51	69.30	69.51	69.93	69.19	68.77
MathVista	71.10	71.00	70.70	70.70	70.60	69.60	70.90	70.70	69.90
We-Math	67.01	66.03	65.29	66.44	65.92	66.67	66.15	65.80	66.72
MathVerse	47.26	47.18	46.02	47.92	46.73	46.83	46.83	47.31	47.39
MathVision	28.27	28.06	26.25	27.91	27.16	27.62	27.25	28.36	28.03
Avg.	56.46	56.29	55.53	56.50	55.94	56.05	56.21	56.27	56.16
<i>Avg.(Std.)</i>									
Avg.(Std.)	56.53(±0.21)	56.12(±0.18)	55.75(±0.19)	56.24(±0.23)	56.04(±0.09)	56.51(±0.43)	56.06(±0.24)	56.20(±0.18)	56.08(±0.09)

Table 4: Preliminary comparison of different token selection strategies in the VLM setting across three independent training runs. We compare full-token training, entropy-based Top- k % selection, and random token selection under multiple retention ratios, and additionally report the mean and standard deviation across runs. Overall, entropy-based selection provides only limited and unstable gains in the VLM setting, remaining close to random selection and generally failing to consistently surpass the full-token baseline.

full-token baseline. A representative phenomenon is that the best entropy-based variant remains only on par with, or slightly below, the full-token setting, while the gap between Top- k % and Random- k % remains small across different retention ratios. This indicates that, for multimodal reasoning, token entropy alone is not a reliable proxy for visually critical positions.

Overall, these preliminary results reveal a clear discrepancy between the two settings: while entropy-based token selection is effective in pure language reasoning and follows the 80/20 rule (Wang et al., 2026), its advantage becomes much weaker in visual reasoning, where token importance must additionally reflect visual dependency rather than uncertainty alone.

C Implementation Details of VEPO

Hyperparameter	Value
Global batch size	128
Rollout batch size	512
Rollout number	12
Rollout temperature	1.0
Freeze vision tower	True
KL loss	False
Learning rate	1.0×10^{-6}
Weight decay	1.0×10^{-2}
Noisy augmentation	Gaussian
Noise step	500

Table 5: Main training configurations for VEPO.

The main training configurations are summarized in Table 5. We use a global batch size of 128 and a rollout batch size of 512, with 12 rollouts per prompt at a rollout temperature of 1.0. The policy model is optimized using a learning rate

Model / Dataset	Geo3K	MMK12	HalluBench	MathVista	We-Math	MathVerse	MathVision	Avg.
<i>Visual-focused RL Fine-tuning Methods (3B)</i>								
NoisyRollout-3B (Liu et al., 2025b)	42.26	58.12	63.85	62.20	61.78	40.94	25.07	50.60
PAPD-DAPD-3B (Wang et al., 2025d)	42.60	<u>58.65</u>	<u>62.32</u>	<u>63.25</u>	<u>62.47</u>	40.65	25.00	50.71
VPPO-3B (Huang et al., 2025)	42.15	58.16	58.60	64.20	62.30	<u>41.56</u>	24.51	50.21
R1-ShareVL (Yao et al., 2026)	39.93	56.26	61.41	60.30	61.26	40.40	25.23	49.26
<i>Our Method</i>								
Qwen2.5-VL-3B-Instruct (Bai et al., 2025)	26.96	45.35	60.88	56.50	53.97	32.90	21.78	42.62
+ GRPO	<u>43.43</u>	57.37	59.52	62.16	61.15	40.95	<u>25.57</u>	50.02
+ GRPO (Top 20% high entropy)	38.28	56.81	59.20	61.40	59.66	37.74	24.18	48.18
+ GRPO (Top 40% high entropy)	40.77	54.52	59.73	62.60	59.77	38.78	24.05	48.60
+ VEPO*	44.32 (± 1.51)	59.44 (± 1.11)	60.50 (± 0.06)	63.07 (± 1.52)	62.59 (± 0.53)	43.16 (± 0.25)	25.93 (± 0.39)	51.29 (± 0.13)
Δ vs. GRPO (Top 20% high entropy)	(+6.04)	(+2.63)	(+1.30)	(+1.67)	(+2.93)	(+5.42)	(+1.75)	(+3.11)

Table 6: Additional results on the 3B model. Consistent with our main findings, VEPO (bottom) outperforms all corresponding baselines at this scale. Results for models marked with * are reported as mean ($\pm$ standard deviation) over multiple training runs with different random seeds. Best and second-best results are determined only among the visual-focused RL fine-tuning methods and our methods, and are highlighted in **bold** and underlined, respectively.

α / Dataset	Geo3K	MMK12	HalluBench	MathVista	We-Math	MathVerse	MathVision	Avg.
$k = 0.2$								
$\alpha = 0.1$	47.09	67.53	68.77	69.30	66.44	48.32	28.13	56.51
$\alpha = 0.3$	48.25	67.78	68.24	71.30	67.24	48.43	27.20	56.92
$\alpha = 0.5$	49.42	68.65	68.45	70.80	67.99	48.91	27.27	57.36
$\alpha = 0.7$	51.75	69.31	71.12	72.37	69.62	49.15	28.42	58.82
$\alpha = 0.9$	50.08	67.53	68.45	71.30	67.36	50.10	27.86	57.53
$k = 0.3$								
$\alpha = 0.1$	49.25	66.17	68.03	70.00	65.46	48.91	26.91	56.39
$\alpha = 0.3$	50.75	68.65	70.70	71.70	68.10	49.82	28.22	58.28
$\alpha = 0.5$	50.92	68.16	71.29	70.30	68.28	49.90	28.12	58.14
$\alpha = 0.7$	52.08	68.40	69.51	71.50	68.16	48.55	27.70	57.99
$\alpha = 0.9$	48.92	68.03	67.51	69.80	67.99	48.20	27.50	56.85
$k = 0.4$								
$\alpha = 0.1$	47.42	66.42	67.30	68.90	67.07	44.31	28.79	55.74
$\alpha = 0.3$	48.25	66.91	69.30	64.50	67.36	48.15	25.93	55.77
$\alpha = 0.5$	48.25	64.81	67.19	70.30	68.39	46.27	26.61	55.97
$\alpha = 0.7$	51.25	68.77	68.56	69.60	67.76	47.66	28.06	57.38
$\alpha = 0.9$	48.92	67.66	67.82	68.90	66.55	46.98	27.07	56.27

Table 7: Detailed numerical results for the ablation on the token selection ratio k and the balancing coefficient α . The table reports the performance on each benchmark together with the overall average for every (k, α) configuration. The default VEPO setting ($k=0.2, \alpha=0.7$) is reported as the multi-seed average, consistent with the main results.

of 1.0×10^{-6} and a weight decay of 1.0×10^{-2} . To improve efficiency and stability, we freeze the vision tower, disable KL loss, and adopt tensor parallelism with size 4 on 8 H800 GPUs. For noisy visual augmentation, we apply Gaussian perturbation at diffusion step 500 and use a sigmoid decay schedule with coefficient 30 and midpoint step 40. For automatic evaluation, we use an external judge model (gpt-4o-mini) to parse model outputs and score final answers.

Dataset and Benchmark Details. For in-domain evaluation, we report results on the validation splits of Geometry3K (Lu et al., 2021) and MMK12 (Meng et al., 2025), which measure performance within the training distribution. For out-of-domain evaluation, we use five public benchmarks to assess

generalization beyond the training domains: four visual mathematical reasoning benchmarks, including MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024b), MathVista (Lu et al., 2024), and We-Math (Qiao et al., 2024), together with the visual perception benchmark HalluBench (Guan et al., 2024).

Baseline. We compare our method against four groups of baselines: (1) closed-source multi-modal models, including GPT-4o (OpenAI et al., 2024), o1 (OpenAI et al., 2026), Claude-3.7-Sonnet (Anthropic, 2025a), and Gemini-2.0-Flash (Gemini Team, Google, 2025); (2) open-source large-scale vision-language models, including QVQ-

Variant	Geo3K	MMK12	HalluBench	MathVista	We-Math	MathVerse	MathVision	Avg.
Qwen2.5-VL-7B-Instruct	35.94	53.66	63.51	66.87	62.13	43.27	25.10	50.07
GRPO	49.42	65.18	70.45	71.60	67.18	48.53	27.40	57.11
Ours w/ Masked Image	52.58	68.65	71.19	71.60	68.65	49.21	27.93	58.54
Ours w/ Gaussian Noise	51.75	69.31	71.12	72.37	69.62	49.15	28.42	58.82

Table 8: Ablation on the role of visual perturbation in VEPO. We compare the base model, standard GRPO, our method with masked image input, and the full VEPO variant with Gaussian perturbation. The Gaussian variant is reported as the multi-seed average, consistent with the main results.

72B-Preview (Team, 2024), Qwen2.5-VL-32B³ (Bai et al., 2025), Qwen2.5-VL-72B (Bai et al., 2025), InternVL2.5-38B (Chen et al., 2025b), and InternVL2.5-78B (Chen et al., 2025b); (3) visual-focused RL fine-tuning methods, including **VPPO** (Huang et al., 2025), **PAPO-DAPO** (Wang et al., 2025d), **R1-ShareVL** (Yao et al., 2026), and **Noisy-Rollout** (Liu et al., 2025b), which are trained on the same 4.2K mixed training set for a fair comparison; and (4) two baselines that isolate the effect of entropy-based token selection: the top-entropy variant **80/20 Rule** (Wang et al., 2026) and the full-token **GRPO** (Shao et al., 2024) baseline.

D Results on Qwen2.5-VL-3B-Instruct

VEPO achieves the best average of **51.29**, surpassing the strongest visual-focused RL baseline PAPO-DAPO-3B (50.71, based on DAPO (Yu et al., 2026)) and exceeding VPPO-3B by **+1.08**, also improving over the GRPO (Top 20% high entropy) baseline by **+3.11**, while ranking first on five of seven benchmarks.

E More Ablation Studies and Analysis

E.1 Detailed Ablation Results for α and k

The coefficient α balances the contribution of the distribution-shift signal JSD and the entropy-gap signal $|\Delta H|$, while k controls the fraction of top-ranked tokens selected for policy updates in each response. We vary α in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and k in $\{0.2, 0.3, 0.4\}$ to examine the impact of different scoring preferences and selection granularities. The extreme cases $\alpha = 0$ and $\alpha = 1$ are excluded here, as they correspond to component-wise variants and are analyzed separately in our scoring component ablation. Table 7 reports the detailed numerical results of the hyperparameter ablation on the token selection ratio k and the balancing

coefficient α . These results complement the trends summarized in Figure 4 in the main text.

E.2 Ablations on Visual Perturbation

We further analyze the role of visual perturbation by comparing two VEPO variants: one using masked image input and the other using Gaussian-noise perturbation. As shown in Table 8, the Gaussian variant achieves the best overall average performance (58.82 vs. 58.54) and outperforms the masked-image variant on four of the seven benchmarks, including MMK12, MathVista, We-Math, and MathVision. The gains are especially consistent on benchmarks that require stronger visual perception and fine-grained grounding, such as MathVision.

This pattern suggests that the form of perturbation plays an important role in token selection. Compared with masked image input, Gaussian noise perturbs the visual evidence while largely preserving the global image structure. As a result, the discrepancy between the original-image and perturbed-image predictions more faithfully reflects fine-grained visual dependency, rather than a broader degradation caused by an uninformative visual input. In this sense, Gaussian perturbation provides a more informative counterfactual signal for identifying visually dependent tokens.

Overall, this comparison supports the design choice in VEPO: the benefit comes not merely from contrasting two visual conditions, but from using a perturbation that preserves enough structure to reveal meaningful token-level visual sensitivity.

F Templates

System Prompt:

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think>` `</think>` tags. The final answer MUST BE put in `\boxed{}`.

User Input:

{question}

³Qwen2.5-VL-32B is a later release in the same series (March 2025) and is not covered by the technical report, which documents only the 3B, 7B, and 72B variants.

Model / Dataset	Geo3K	MMK12	HalluBench	MathVista	We-Math	MathVerse	MathVision	Avg.	Avg. (w/o Geo3K, MMK12)
GRPO	49.42	65.18	70.45	71.60	67.18	48.53	27.40	57.11	57.03
	50.58	66.44	68.24	69.90	68.10	48.02	27.70	57.00	56.39
	49.08	67.29	68.98	70.40	67.99	47.31	27.66	56.96	56.47
Avg. (Std.)	49.69 (± 0.79)	66.30 (± 1.06)	69.22 (± 1.12)	70.63 (± 0.87)	67.76 (± 0.50)	47.95 (± 0.61)	27.59 (± 0.16)	57.02 (± 0.08)	56.63 (± 0.35)
PAPO-DAPO-7B	50.08	66.05	72.24	73.10	68.69	48.07	26.23	57.78	57.66
	49.75	66.67	71.08	72.30	67.99	48.45	27.30	57.65	57.42
	50.42	66.79	71.19	72.10	68.79	48.20	27.51	57.86	57.56
Avg. (Std.)	50.08 (± 0.34)	66.50 (± 0.40)	71.50 (± 0.64)	72.50 (± 0.53)	68.49 (± 0.44)	48.24 (± 0.19)	27.01 (± 0.69)	57.76 (± 0.11)	57.55 (± 0.12)
VPPO-7B	50.42	69.02	70.14	72.10	69.43	48.21	28.19	58.22	57.61
	50.58	68.65	70.03	71.10	69.43	48.25	27.89	57.99	57.34
	50.58	66.67	71.40	71.30	67.93	49.92	27.80	57.94	57.67
Avg. (Std.)	50.53 (± 0.09)	68.11 (± 1.26)	70.52 (± 0.76)	71.50 (± 0.53)	68.93 (± 0.87)	48.79 (± 0.98)	27.96 (± 0.20)	58.05 (± 0.15)	57.54 (± 0.18)
VEPO-7B	51.58	69.64	71.71	72.00	69.54	48.93	28.31	58.82	58.10
	52.58	68.77	71.19	72.60	69.60	48.10	28.29	58.73	57.96
	51.08	69.52	70.45	72.50	69.71	50.43	28.65	58.91	58.35
Avg. (Std.)	51.75 (± 0.76)	69.31 (± 0.47)	71.12 (± 0.63)	72.37 (± 0.32)	69.62 (± 0.09)	49.15 (± 1.18)	28.42 (± 0.20)	58.82 (± 0.09)	58.14 (± 0.20)

Table 9: Multi-seed results for the 7B setting. Each method is trained using three random seeds, with the mean and standard deviation reported in the final row of each block.

Model / Dataset	Geo3K	MMK12	HalluBench	MathVista	We-Math	MathVerse	MathVision	Avg.	Avg. (w/o Geo3K, MMK12)
PAPO-DAPO-3B	42.76	58.85	62.36	63.56	62.30	40.79	25.10	50.82	50.82
	42.60	58.12	61.93	63.20	62.64	41.02	25.33	50.69	50.82
	42.43	58.98	62.67	63.00	62.47	40.15	24.57	50.61	50.57
Avg. (Std.)	42.60 (± 0.17)	58.65 (± 0.46)	62.32 (± 0.37)	63.25 (± 0.28)	62.47 (± 0.17)	40.65 (± 0.45)	25.00 (± 0.39)	50.71 (± 0.11)	50.74 (± 0.14)
VPPO-3B	42.10	58.24	58.46	64.20	62.64	41.57	24.54	50.25	50.28
	42.43	57.74	58.89	63.90	61.95	41.73	24.01	50.09	50.10
	41.93	58.49	58.46	64.50	62.30	41.37	24.97	50.29	50.32
Avg. (Std.)	42.15 (± 0.25)	58.16 (± 0.38)	58.60 (± 0.25)	64.20 (± 0.30)	62.30 (± 0.35)	41.56 (± 0.18)	24.51 (± 0.48)	50.21 (± 0.11)	50.23 (± 0.12)
VEPO-3B	44.93	58.86	60.46	62.80	62.70	43.32	26.25	51.33	51.11
	42.60	58.74	60.57	64.70	62.01	43.30	26.05	51.14	51.33
	45.42	60.72	60.46	61.70	63.05	42.87	25.49	51.39	50.71
Avg. (Std.)	44.32 (± 1.51)	59.44 (± 1.11)	60.50 (± 0.06)	63.07 (± 1.52)	62.59 (± 0.53)	43.16 (± 0.25)	25.93 (± 0.39)	51.29 (± 0.13)	51.05 (± 0.31)

Table 10: Multi-seed results for the 3B setting. Each method is trained using three random seeds, with the mean and standard deviation reported in the final row of each block.

Model / Dataset	Geo3K	MMK12	HalluBench	MathVista	We-Math	MathVerse	MathVision	Avg. $ D $ over Steps (% of VEPO Tokens)	Avg.
Turn 1	48.59	66.67	69.61	71.60	68.16	48.71	27.83	15.2	57.31
Turn 2	48.75	66.54	70.98	71.20	67.76	48.05	28.09	13.7	57.34
Turn 3	49.25	67.16	69.61	71.00	67.70	48.12	27.93	14.8	57.25
Avg. (Std.)	48.86 (± 0.34)	66.79 (± 0.33)	70.07 (± 0.79)	71.27 (± 0.31)	67.87 (± 0.25)	48.29 (± 0.36)	27.95 (± 0.13)	14.57	57.30 (± 0.05)
VEPO	51.75 (± 0.76)	69.31 (± 0.47)	71.12 (± 0.63)	72.37 (± 0.32)	69.62 (± 0.09)	49.15 (± 1.18)	28.42 (± 0.20)	–	58.82 (± 0.09)

Table 11: Downstream performance under targeted ablation of high-JSD, low-entropy tokens. Each turn removes the high-JSD, low-entropy subset from the VEPO selection. We report three independent runs together with the mean and standard deviation. VEPO results are the multi-seed averages from the main experiments.

G Theoretical Interpretation via Aleatoric–Epistemic Uncertainty Decomposition

In this appendix, we provide an information-theoretic interpretation of VEPO’s joint vision-entropy token selection design (Eq. 7 in the main paper). In particular, we show that the two visual signals: Jensen–Shannon Divergence (JSD, (Menéndez et al., 1997)) and the absolute entropy gap $|\Delta H_t|$ can be interpreted as capturing two complementary aspects of predictive change under visual perturbation. We further show that the multiplicative aggregation in g_t admits a soft noisy-style OR (Pearl, 2014).

G.1 Preliminaries

Setup. Recall from Eq. (1) that VEPO performs two forward passes at each token position t :

$$\begin{aligned} P_t(\cdot) &= \pi_\theta(\cdot \mid q, I, o_{<t}), \\ Q_t(\cdot) &= \pi_\theta(\cdot \mid q, I', o_{<t}), \end{aligned} \quad (12)$$

where I is the original image and I' is its noise-perturbed counterpart.

Visual condition. We introduce a binary variable $\Theta_v \in \{I, I'\}$ indicating whether the next-token prediction is conditioned on the original or perturbed image. Assuming a uniform prior $p(\Theta_v=I) = p(\Theta_v=I') = \frac{1}{2}$, the marginal predictive distribution of the next token Y_t is

$$\begin{aligned} M_t(\cdot) &= \mathbb{E}_{\Theta_v}[\pi_\theta(\cdot \mid q, \Theta_v, o_{<t})] \\ &= \frac{1}{2}(P_t + Q_t), \end{aligned} \quad (13)$$

which coincides with the mixture distribution used in the definition of JSD (Eq. 2).

Uncertainty decomposition. Following the standard aleatoric–epistemic decomposition (Depeweg et al., 2018; Houlisby et al., 2011; Kendall and Gal, 2017), the predictive uncertainty of Y_t under this visual-condition mixture can be decomposed as

$$\underbrace{H(Y_t | q, o_{<t})}_{\text{Total Uncertainty (TU)}_t} = \underbrace{I(Y_t; \Theta_v | q, o_{<t})}_{\text{Epistemic Component (EU)}_t} + \underbrace{\mathbb{E}_{\Theta_v}[H(Y_t | q, \Theta_v, o_{<t})]}_{\text{Aleatoric Uncertainty (AU)}_t}. \quad (14)$$

Here, the aleatoric term captures the residual uncertainty within each fixed visual condition, while the epistemic term captures the reducible disagreement across visual conditions.

G.2 Proposition 1: JSD as a Visual-Condition Epistemic Component

Proposition G.1 (JSD as conditional mutual information). Under the setup of §G.1, the Jensen–Shannon Divergence between P_t and Q_t is exactly the conditional mutual information between the next token and the visual condition:

$$\text{JSD}(P_t \| Q_t) = I(Y_t; \Theta_v | q, o_{<t}). \quad (15)$$

Proof. By the definition of mutual information,

$$\begin{aligned} & I(Y_t; \Theta_v | q, o_{<t}) \\ &= H(Y_t | q, o_{<t}) - H(Y_t | q, \Theta_v, o_{<t}) \\ &= H(M_t) - \mathbb{E}_{\Theta_v}[H(Y_t | q, \Theta_v, o_{<t})] \\ &= H(M_t) - \frac{1}{2}[H(P_t) + H(Q_t)]. \end{aligned} \quad (16)$$

Meanwhile, the Jensen–Shannon Divergence with uniform mixing weights (Lin, 1991) satisfies

$$\text{JSD}(P_t \| Q_t) = H(M_t) - \frac{1}{2}[H(P_t) + H(Q_t)], \quad (17)$$

which matches Eq. (16). $\square$

Remark. Proposition G.1 shows that JSD is exactly the mutual information between the next token and the visual condition. In this sense, JSD quantifies how strongly the next-token prediction depends on the visual input.

G.3 Proposition 2: $|\Delta H_t|$ as Condition-Wise Aleatoric Change

Proposition G.2 ($|\Delta H_t|$ as aleatoric change). The absolute entropy gap $|\Delta H_t|$ defined in Eq. (6) measures the magnitude of the change in conditional

predictive entropy induced by visual perturbation:

$$|\Delta H_t| = |H(Y_t | q, \Theta_v=I', o_{<t}) - H(Y_t | q, \Theta_v=I, o_{<t})|. \quad (18)$$

Proof. Under each fixed visual condition, the conditional predictive entropy is

$$\begin{aligned} H(Y_t | q, \Theta_v=I, o_{<t}) &= H(P_t), \\ H(Y_t | q, \Theta_v=I', o_{<t}) &= H(Q_t). \end{aligned} \quad (19)$$

Therefore, by Eq. (6),

$$\Delta H_t = H(Q_t) - H(P_t), \quad (20)$$

and its absolute value equals the magnitude of the condition-wise entropy change. $\square$

Remark. Unlike the aleatoric term in Eq. (14), which averages uncertainty across visual conditions, $|\Delta H_t|$ measures how much the predictive entropy changes between them. A large $|\Delta H_t|$ therefore indicates that the token is sensitive to the visual input, whether the perturbation makes the model more confident or less confident.

G.4 Proposition 3: Non-Equivalence and Complementarity of JSD and $|\Delta H_t|$

Proposition G.3 (Non-equivalence and complementarity). The two visual signals $\text{JSD}(P_t \| Q_t)$ and $|\Delta H_t| = |H(Q_t) - H(P_t)|$ capture complementary aspects of the predictive shift between P_t and Q_t . In particular, neither signal determines the other alone: (i) $|\Delta H_t|$ may vanish while $\text{JSD}(P_t \| Q_t)$ remains strictly positive, and (ii) distinct pairs of distributions may share the same $|\Delta H_t|$ while exhibiting different JSD values.

Proof by construction. We provide two examples on a 3-symbol vocabulary $\mathcal{V} = \{v_1, v_2, v_3\}$.

Case (a): Zero entropy gap but positive JSD.

Let

$$P_t = (1, 0, 0), \quad Q_t = (0, 1, 0).$$

Then both distributions are deterministic, so

$$\begin{aligned} H(P_t) &= H(Q_t) = 0, \\ |\Delta H_t| &= |H(Q_t) - H(P_t)| = 0. \end{aligned}$$

However, P_t and Q_t place all probability mass on different tokens, so

$$\text{JSD}(P_t \| Q_t) = \log 2 > 0.$$

Case (b): Same entropy gap but different JSD.
Consider the two pairs:

$$P_t^{(1)} = (1, 0, 0), \quad Q_t^{(1)} = (\frac{1}{2}, \frac{1}{2}, 0),$$

and

$$P_t^{(2)} = (0, 0, 1), \quad Q_t^{(2)} = (\frac{1}{2}, \frac{1}{2}, 0).$$

For both pairs,

$$\begin{aligned} H(P_t^{(1)}) &= H(P_t^{(2)}) = 0, \\ H(Q_t^{(1)}) &= H(Q_t^{(2)}) = \log 2. \end{aligned}$$

Hence,

$$|\Delta H_t^{(1)}| = |\Delta H_t^{(2)}| = \log 2.$$

However, their Jensen–Shannon divergences differ:

$$\begin{aligned} \text{JSD}(P_t^{(1)} \| Q_t^{(1)}) &= \frac{3}{4} \log \frac{4}{3}, \\ \text{JSD}(P_t^{(2)} \| Q_t^{(2)}) &= \log 2. \end{aligned}$$

Thus, the same entropy gap may correspond to substantially different distributional disagreement.

Taken together, these examples show that $\text{JSD}(P_t \| Q_t)$ and $|\Delta H_t|$ are non-equivalent and complementary: one may vanish while the other is positive, and even when the entropy gap is fixed, JSD can still vary substantially. $\square$

Remark. Proposition G.3 clarifies the role of the two visual signals in VEPO. JSD captures disagreement between the predictive distributions under different visual conditions, whereas $|\Delta H_t|$ measures how much the corresponding predictive uncertainty changes. Because these signals encode different aspects of the predictive shift, relying on only one of them may fail to recover all visually sensitive tokens. This is consistent with the ablation results in Table 2, where removing either signal degrades performance.

G.5 Proposition 4: g_t as a Soft Noisy-OR-Style Aggregation

Proposition G.4 (Soft noisy-OR aggregation). The joint visual-dependency score:

$$g_t = 1 - (1 - \hat{j}_t)^\alpha (1 - |\widehat{\Delta H}_t|)^{1-\alpha} \quad (21)$$

admits a noisy-OR-style interpretation as a soft union over two complementary visual-sensitivity cues.

Proof. For two independent binary events A and B with marginal probabilities p_A and p_B , the classical noisy-OR-style takes the form:

$$P(A \cup B) = 1 - (1 - p_A)(1 - p_B). \quad (22)$$

VEPO’s score in Eq. (21) has the same complement-product structure, while introducing weights α and $1 - \alpha$ to control the relative emphasis placed on JSD and $|\Delta H_t|$. As $\alpha \rightarrow 1$, $g_t \rightarrow \hat{j}_t$; as $\alpha \rightarrow 0$, $g_t \rightarrow |\widehat{\Delta H}_t|$. Thus, g_t can be viewed as a soft union operator that prioritizes tokens that are salient under either visual cue, while still allowing one cue to dominate when appropriate. $\square$

Remark. Proposition G.4 provides a probabilistic intuition for the multiplicative aggregation in VEPO. Unlike additive fusion, $\alpha \hat{j}_t + (1 - \alpha) |\widehat{\Delta H}_t|$, the complement-product form saturates more naturally when either cue is strong, making it better suited for retaining tokens that spike along one visual-sensitivity axis but not necessarily the other. This interpretation is consistent with our empirical observation (Table 2) that replacing the multiplicative aggregation with additive fusion degrades performance.

G.6 Putting It Together: The Full Scoring Function

Combining Propositions G.1–G.4, VEPO’s per-token score

$$\begin{aligned} c_t &= g_t \cdot \hat{h}_t \\ &= \left[1 - (1 - \hat{j}_t)^\alpha (1 - |\widehat{\Delta H}_t|)^{1-\alpha} \right] \hat{h}_t. \end{aligned} \quad (23)$$

Here, the bracketed term acts as a soft union over the two complementary visual cues, while $\hat{h}_t$ measures token informativeness. This score admits a three-layer interpretation:

- (i) $\hat{j}_t$ captures visual-condition disagreement through conditional mutual information (Proposition G.1);
- (ii) $|\widehat{\Delta H}_t|$ captures the magnitude of condition-wise predictive-entropy change (Proposition G.2);
- (iii) g_t combines these complementary cues through a noisy-OR-style soft union, while multiplication by $\hat{h}_t$ further prioritizes tokens that are also informative under the original (unperturbed) distribution.

This interpretation helps explain why VEPO’s joint vision-entropy token selection strategy is effective at identifying visually grounded and highly informative tokens.

H LLM Usage Statement

We used Large Language Models (LLMs) as auxiliary tools to assist with the writing process. They were used solely to polish the language and improve readability, with no influence over the research design, experimental implementation or analysis. We conceived and executed all methodological contributions, experiments, and conclusions independently.

I Algorithm

Algorithm 1 summarizes the VEPO pipeline. It first generates responses with standard rollout, then computes token-level visual-dependency signals by comparing model predictions under the original and perturbed images. Based on the resulting token scores, only the top- k fraction of tokens is retained for policy gradient updates.

Algorithm 1 VEPO: Vision-Entropy Token Selection for Policy Optimization

Require: Policy model π_θ ; prompt set $\mathcal{D}$; original image I ; noise step τ ; score weight α ; retention ratio k

Ensure: Updated policy π_θ

```

1: for each training iteration do
2:   // Stage 1: Rollout with original image
3:   Sample responses  $\{o^{(k)}\}_{k=1}^n \sim \pi_\theta(\cdot \mid q, I)$ 
4:   Compute reward  $r^{(k)}$  and group-relative advantage  $A^{(k)}$ 
5:   // Stage 2: Counterfactual forward pass
6:   Construct perturbed image
7:    $I' \leftarrow \sqrt{\alpha_\tau} I + \sqrt{1 - \alpha_\tau} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
8:   for each response  $o = (o_1, \dots, o_T)$  do
9:     for each token position  $t = 1, \dots, T$  do
10:       $P_t \leftarrow \pi_\theta(\cdot \mid q, I, o_{<t})$  // original image forward
11:       $Q_t \leftarrow \pi_\theta(\cdot \mid q, I', o_{<t})$  // perturbed image forward
12:    end for
13:  end for
14:  // Stage 3: Compute token-level visual dependency signals
15:  for each token position  $t = 1, \dots, T$  do
16:     $M_t \leftarrow \frac{1}{2}(P_t + Q_t)$ 
17:     $\text{JSD}_t \leftarrow \frac{1}{2}\text{KL}(P_t \parallel M_t) + \frac{1}{2}\text{KL}(Q_t \parallel M_t)$ 
18:     $H_t \leftarrow -\sum_v P_t(v) \log P_t(v)$ 
19:     $H_t^{\text{mask}} \leftarrow -\sum_v Q_t(v) \log Q_t(v)$ 
20:     $\Delta H_t \leftarrow H_t^{\text{mask}} - H_t$ 
21:  end for
22:  // Stage 4: Score computation and token selection
23:   $\hat{j}_t, |\widehat{\Delta H}_t|, \hat{h}_t \leftarrow \text{MinMaxNorm}(\text{JSD}_t, |\Delta H_t|, H_t)$  // per-response
24:  for each token position  $t = 1, \dots, T$  do
25:     $g_t \leftarrow 1 - (1 - \hat{j}_t)^\alpha \cdot (1 - |\widehat{\Delta H}_t|)^{1-\alpha}$  // joint visual-dependency signal
26:     $c_t \leftarrow g_t \cdot \hat{h}_t$  // entropy-gated score
27:  end for
28:   $\mathcal{C}_t \leftarrow 1[c_t \in \text{Top-}[k \cdot T](\{c_\tau\}_{\tau=1}^T)]$  // binary mask
29:  // Stage 5: Masked policy gradient update
30:   $\mathcal{L}_{\text{VEPO}} \leftarrow \frac{\sum_{t=1}^T \ell_t \cdot m_t \cdot \mathcal{C}_t}{\sum_{t=1}^T m_t \cdot \mathcal{C}_t}$ 
31:  Update  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{VEPO}}$ 
32: end for

```

J Qualitative Cases

Qualitative Case 1: VEPO vs. Top-Entropy

Question: The running result of the following program is ____.

Ground truth: $\frac{19}{3}$

```
a=2
b=3
c=4
a=b
b=c+2
c=b+4
d=(a+b+c)/3
PRINT "d=";d
END
```

Legend. Both VEPO-only Entropy-only

Generated response.

< think >

The reasoning process is as follows:

1. Initialize the variables: $a = 2$, $\quad b = 3$, $c = 4$
2. Assign b to a : $a = b = 3$
3. Assign $c + 2$ to b : $b = c + 2 = 4 + 2 = 6$
4. Assign $b + 4$ to c : $c = b + 4 = 6 + 4 = 10$
5. Calculate d as the average of a , b , and c : $d = \frac{a + b + c}{3} = \frac{3+6+10}{3} = \frac{19}{3}$
6. Print the value of d : $d = \frac{19}{3}$

</ think > $\boxed{\frac{19}{3}}$

Representative Token Selections.

Both: process Initialize variables Assign Assign Calculate Print

VEPO-only: a c + b b + c as the the

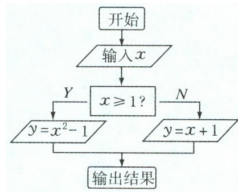
Entropy-only: think reasoning :\\n\\n quad):\\n [\\n]\\n\\n) frac boxed

Figure 6: A qualitative comparison between VEPO and Top-Entropy on a visually grounded program-tracing example. The upper panel shows the program image and generated response, while the lower panel groups representative token selections into shared, VEPO-only, and entropy-only categories.

Qualitative Case 2: VEPO vs. Top-Entropy

Question: As shown in the figure, according to the given program, if the input is $x = \sqrt{3}$, then the output result is _____.

Ground truth: 2



Legend. Both Both VEPO-only Entropy-only

Generated response.

<think>

The reasoning process is as follows :

Since the input is $x = \sqrt{3}$, we check the condition $x \geq 1$. Here, $\sqrt{3} \approx 1.732$, which is greater than 1. Therefore, the condition $x \geq 1$ is true.

According to the flowchart, when $x \geq 1$, the value of y is given by $y = x^2 - 1$. So, we substitute $x = \sqrt{3}$ into the equation :

$$y = (\sqrt{3})^2 - 1 = 3 - 1 = 2$$

Thus, the output result is 2.</think> \boxed{2}

Representative Token Selections.

Both: reasoning process input check condition flowchart value output result

VEPO-only: greater than when given equation

Entropy-only: follows q which we think

Figure 7: Another qualitative comparison between VEPO and Top-Entropy on a visually grounded program-tracing example.