Dual-Uncertainty Guided Policy Learning for Multimodal Reasoning

Rui Liu[1,2], **Dian Yu**[1], **Tong Zheng**[2], **Runpeng Dai**[3], **Zongxia Li**[2], **Wenhao Yu**[1], **Zhenwen Liang**[1], **Linfeng Song**[4*], **Haitao Mi**[1], **Pratap Tokekar**[2], **Dong Yu**[1]

[1]Tencent Hunyuan [2]University of Maryland, College Park
[3]University of North Carolina at Chapel Hill [4]Google

Code: http://github.com/ruiiu/DUPL

Abstract

Reinforcement learning with verifiable rewards (RLVR) has advanced reasoning capabilities in multimodal large language models. However, existing methods typically treat visual inputs as deterministic, overlooking the perceptual ambiguity inherent to the visual modality. Consequently, they fail to distinguish whether a model's uncertainty stems from complex reasoning or ambiguous perception, preventing the targeted allocation of exploration or learning signals. To address this gap, we introduce **DUPL**, a dual-uncertainty guided policy learning approach for multimodal RLVR that quantifies and leverages both perceptual uncertainty (via symmetric KL divergence) and output uncertainty (via policy entropy) to guide policy updates. By establishing an uncertainty-driven feedback loop and employing a dynamic branch prioritization mechanism, DUPL recalibrates the policy advantage to focus learning on states with high perceptual or decisional ambiguity, enabling effective targeted exploration beyond passive data augmentation. Evaluated on diverse multimodal reasoning benchmarks spanning mathematical and general domains, DUPL achieves solid gains. It improves Qwen2.5-VL accuracy by up to **12.3%** (3B) and **7.9%** (7B), and Qwen3-VL-Instruct by up to **10.7%** (4B) and **12.4%** (8B), consistently outperforming GRPO, while seamlessly generalizing to alternative algorithms (DAPO, **+6.5%** avg) and architectures (LLaVA-OneVision-1.5, **+4.7%** avg). These results demonstrate that DUPL is an effective and generalizable approach for multimodal RLVR.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has substantially improved the reasoning abilities of large language models (LLMs) by optimizing against ground-truth answers (Luong et al., 2024; Lambert et al., 2024; Guo et al., 2025; Su et al., 2025; Zheng et al., 2025a). However, this outcome-centric approach often suppresses trajectories with valid intermediate reasoning steps that conclude with incorrect final answers, restricting exploration and yielding brittle policies (Dai et al., 2025). While text-only strategies mitigate this via uncertainty-aware objectives (Cheng et al., 2025), diversity-promoting rewards (Li et al., 2025a), Pass@k rewards (Chen et al., 2025b; Walder and Karkhanis, 2025), intermediate feedback (Setlur et al., 2024), and entropy regularization (Cheng et al., 2025; Zhang et al., 2025; Cui et al., 2025; Wang et al., 2025a), these mechanisms operate strictly within the text or action space. As the paradigm shifts to Multimodal Large Language Models (MLLMs) (Huang et al., 2025; Tan et al., 2025; Peng et al., 2025), where textual reasoning is coupled with complex visual inputs, effective exploration becomes fundamentally more challenging and remains largely underexplored.

Current multimodal RLVR approaches typically treat visual input as a fixed, deterministic condition. This overlooks the inherent ambiguity of the visual modality, where ambiguous objects or multiple valid interpretations may exist. While recent approaches (Liu et al., 2025b; Yao et al., 2025b) attempt to improve robustness via visual noise injection during rollout collection, these methods primarily rely on passive data augmentation that leaves the learning objective unchanged, resulting in undirected exploration rather than targeted exploration toward states of genuine uncertainty. Similarly, while perception-aware optimization improves visual grounding (Wang et al., 2025b), it lacks a mechanism to model or leverage uncertainty to guide policy exploration.

This reveals a fundamental limitation in current multimodal RLVR: exploration strategies are **blind to the source of uncertainty**. By failing to distinguish whether a model's uncertainty stems from complex reasoning or ambiguous perception,

*Work done at Tencent.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 49054–49069
October 24-29, 2026 ©2026 Association for Computational Linguistics

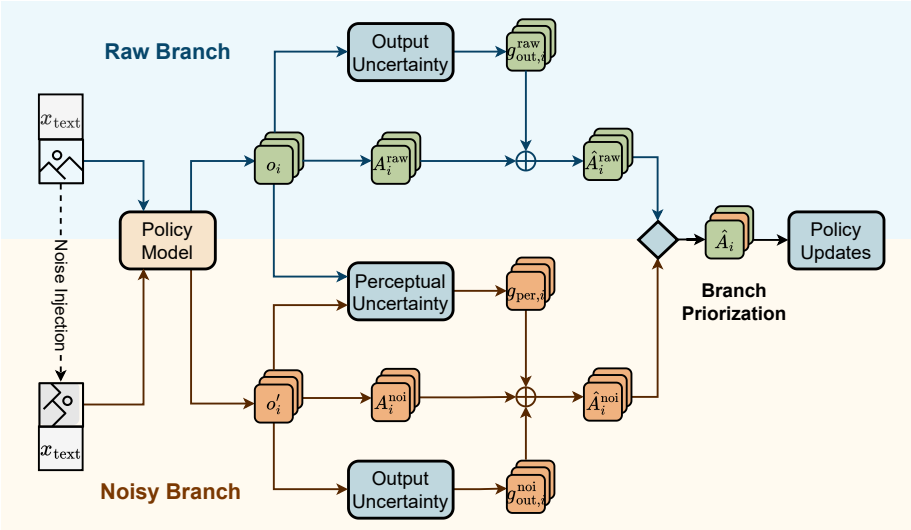


Figure 1: **Overview of DUPL for multimodal reasoning.** DUPL establishes a dual-uncertainty guided feedback loop for policy updates, capturing both perceptual and output uncertainty. The raw branch processes the original input, while the noisy branch processes a perturbed view. Perceptual uncertainty is quantified via token-level symmetric KL divergence between branches, producing the guidance signal g_{per} to shape the advantage. Output uncertainty is measured by policy entropy, generating guidance signals $g_{\text{out}}^{\text{raw}}$ and $g_{\text{out}}^{\text{noi}}$ for the raw and noisy branches, respectively. A dynamic branch prioritization mechanism emphasizes uncertainty-driven exploration early in training and gradually shifts focus to the raw branch as learning stabilizes.

existing methods cannot prioritize learning signals where they are most needed. Consequently, exploration remains inefficiently distributed. What is missing is a policy learning mechanism capable of **decoupling these uncertainty streams** to enable targeted, uncertainty-aware updates.

To address this gap, we introduce **DUPL**, a targeted exploration approach for multimodal RLVR through **D**ual-**U**ncertainty guided **P**olicy **L**earning, as illustrated in Figure 1. Inspired by closed-loop control principles (Hjalmarsson, 2005) and learning under uncertainty (Liu et al., 2023, 2026a,b), DUPL forms an uncertainty-driven feedback loop where model uncertainty acts as a feedback signal to actively regulate policy updates. Rather than encouraging exploration through indiscriminate, passive data augmentation (Liu et al., 2025b; Yao et al., 2025b), DUPL explicitly explores where the model is uncertain, capturing both perceptual and output uncertainty. Specifically, we transform visual perturbations into an **active sensitivity probe**: for each training example, we perform a dual-branch forward pass on the original image and a perturbed view, quantifying **perceptual uncertainty** via the symmetric KL divergence between the induced policy distributions. In parallel, DUPL models **output uncertainty** via the policy's entropy in the action space. By jointly leveraging these two uncertainties as guidance signals to shape advantage for policy updates, DUPL dynamically steers exploration toward perceptually and decision-wise ambiguous states. Furthermore, we employ a dynamic branch prioritization mechanism that emphasizes uncertainty-driven exploration early in training before shifting focus to the original view as learning stabilizes.

To validate DUPL, we evaluate its performance across diverse mathematical and general-domain reasoning benchmarks: MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2023), WeMath (Qiao et al., 2024), HallusionBench (Guan et al., 2024), ChartQA (Masry et al., 2022), LogicVista (Xiao et al., 2024), ChartMuseum (Tang et al., 2026), MMReason (Yao et al., 2025a), and VisuLogic (Xu et al., 2025). When trained on the MMRL30k dataset (Zhu et al., 2025a), DUPL improves Qwen2.5-VL (Bai et al., 2025b) accuracy by up to 12.3% (Avg. 8.7%) at the 3B scale and 7.9% (Avg. 5.8%) at the 7B scale (Table 1), while enhancing Qwen3-VL-Instruct by up to 10.7% (Avg. 5.3%) for the 4B model and up to 12.4% (Avg. 6.3%) for the 8B model (Table 16), consistently outperforming GRPO (Shao et al., 2024). DUPL also demonstrates strong generalization to other RL algorithms: when trained with DAPO (Yu et al., 2026), it improves the 7B base model by an average of 6.5% and surpasses the DAPO baseline

(Table 6). Moreover, applying DUPL to an alternative base model, LLaVA-OneVision-1.5-8B-Instruct (An et al., 2025), yields an average improvement of 4.7% (Table 9). In summary, the core contributions of our work are as follows:

- We identify a core limitation in multimodal RLVR: the inability to separate perceptual ambiguity from reasoning uncertainty, causing inefficient, undirected exploration.
- We introduce DUPL, a policy learning framework that decouples and quantifies perceptual and output uncertainty to recalibrate advantage through an uncertainty-driven feedback loop.
- We propose an active probing mechanism with dynamic branch prioritization, transforming visual perturbations from passive augmentation into a principled signal for targeted, uncertainty-aware optimization.
- Through extensive evaluation, DUPL achieves solid gains. It improves Qwen2.5-VL accuracy by up to **12.3%** (3B) and **7.9%** (7B), and Qwen3-VL-Instruct by up to **10.7%** (4B) and **12.4%** (8B), outperforming GRPO, while seamlessly generalizing to alternative algorithms (DAPO, **+6.5%** avg) and architectures (LLaVA-OneVision-1.5, **+4.7%** avg).

2 Approach

Formally, given a multimodal input $x = (x_{\text{text}}, x_{\text{image}})$, we aim to optimize the MLLM policy network π_θ by maximizing a surrogate objective (e.g., the GRPO objective, detailed in Eq. 3, Appendix A.1).

The core of our approach is to enable targeted exploration in multimodal RLVR through dual-uncertainty guided policy learning. Rather than relying on passive data augmentation, DUPL explicitly explores where the model is uncertain. To this end, we transform visual perturbations into an active sensitivity probe that quantifies perceptual uncertainty. In parallel, we quantify output uncertainty. These two uncertainties serve as active feedback signals: they are integrated into the advantage function to guide policy optimization, incentivizing the model to explore perceptually and decision-wise ambiguous states. Finally, a dynamic branch prioritization schedule modulates this process, steering the model to prioritize exploration in early stages before stabilizing on the raw view for convergence. The full procedure is summarized in Algorithm 1 in Appendix A.2.

Perceptual Uncertainty. Moving beyond the standard use of image augmentation for RL (Yarats et al., 2021; Laskin et al., 2020), we introduce controlled image perturbations and transform them into an active sensitivity probe that quantifies perceptual uncertainty by measuring how the model's output distribution varies under visual transformations. Variations in the model's predictions reflect its sensitivity to plausible perturbations, and therefore identify states worthy of exploration.

Specifically, for each image x_{image} in the training dataset, we create a perturbed counterpart x'_{image} through a stochastic augmentation function $\mathcal{T}$. This function applies a composition of transformations: $x'_{\text{image}} = \mathcal{T}(x_{\text{image}})$, where $\mathcal{T}$ includes random horizontal/vertical flips, rotations, color jittering, and the addition of Gaussian noise. Then we employ a dual-branch forward pass, as shown in Figure 1. The raw branch processes the original input x to produce an output probability distribution $p = \pi_\theta(\cdot|x)$, while the noisy branch uses the perturbed input x' to produce a distribution $q = \pi_\theta(\cdot|x')$. After obtaining the two distributions, we represent the perceptual uncertainty u_{per} as the divergence between them. We measure it using a symmetric KL divergence, which is calculated as the mean of the forward and backward KL divergences, encouraging exploration while maintaining stability:

$$u_{\text{per}} = \frac{1}{2}\big(\mathrm{D}_{\text{KL}}(p||q) + \mathrm{D}_{\text{KL}}(q||p)\big). \quad (1)$$

Output Uncertainty. After modeling the perceptual uncertainty, to promote general policy stochasticity and exploration in the action space, we model the output uncertainty u_{out}, which is based on the token entropy of the policy's output distribution: $u_{\text{out}} = -\sum_{v\in\mathcal{V}} \pi_\theta\left(v \mid x, o_{<t}\right) \log \pi_\theta\left(v \mid x, o_{<t}\right)$, where $\mathcal{V}$ denotes the vocabulary.

Dual-Uncertainty Guided Policy Learning. After obtaining the two uncertainties, we integrate them into the advantage function to guide policy learning. We maintain separate advantage calculations for the raw and noisy branches. For the raw branch, we compute the output guidance signal induced by the output uncertainty, which is defined as: $g_{\text{out}} = \min\Big(\frac{|A|}{\beta_o}, \alpha_o \cdot \mathrm{stopgrad}(u_{\text{out}})\Big)$, where α_o and β_o are scaling factors, $\mathrm{stopgrad}(\cdot)$ is the stop gradient operator, modulating the update

magnitude without affecting gradient propagation.

For the noisy branch, except the output guidance signal, a perceptual guidance signal is incorporated, derived from the measured perceptual uncertainty: $g_{\text{per}} = \min\left(\frac{|A^{\text{noi}}|}{\beta_p}, \alpha_p \cdot \text{stopgrad}(u_{\text{per}})\right)$, where A^{noi} is the advantage for the noisy branch, α_p and β_p are scaling factors.

Therefore, for the raw branch, the uncertainty-guided advantage is calculated as: $\hat{A}^{\text{raw}} = A^{\text{raw}} + g^{\text{raw}}_{\text{out}}$, and for the noisy branch: $\hat{A}^{\text{noi}} = A^{\text{noi}} + g^{\text{noi}}_{\text{out}} + g_{\text{per}}$. When implementing DUPL, we use GRPO's standard estimator to compute the base advantages A^{noi} and A^{raw} (see Eq. 2 in Appendix A.1). Policy updates are then performed using the corresponding uncertainty-guided advantage, depending on the branch selected at each training step.

Dynamic Branch Prioritization. During training, it is crucial to balance the aggressive exploration driven by the noisy branch with the stable learning provided by the raw branch. A policy trained exclusively on the noisy branch may become overly stochastic and fail to converge, while a policy trained solely on the raw branch may not explore enough to find the optimal path. To manage this trade-off, we employ a dynamic branch prioritization strategy. At each training step, we stochastically choose which advantage estimate to use for the policy update. We define p_{noi} as the probability of prioritizing the noisy branch, which is expressed as: $p_{\text{noi}}(s) = \max\left(0, 1 - \frac{s}{s_{\text{total}}}\right)$, where s is the current training step, s_{total} is the total training steps. This probability is decayed over the course of training. Initially, p_{noi} is high to promote broad exploration of the state space. As training progresses, p_{noi} is gradually decreased, causing the optimizer to favor the more stable advantage estimates from the raw branch.

3 Experiments

3.1 Experimental Setup

Implementation Details. We conduct direct RL training on the Qwen2.5-VL-3B and 7B (Bai et al., 2025b) models. The models are trained to generate responses in a structured format, where the reasoning process is enclosed within `<think></think>` tags and the final answer is presented in `\boxed{}`. We train all models on the MMRL30k dataset (Zhu et al., 2025a), which contains around 30K samples. We generate 5 rollouts per input with a rollout batch size of 256. The implementation builds on the framework EasyR1 (Zheng et al., 2025b).

To inject perturbation into images, we apply random horizontal/vertical flips, rotations, color jittering, and the addition of Gaussian noise with zero mean and standard deviation $\sigma = 0.4$. A sensitivity analysis on different noise levels is provided in Section 3.4. For more training details, please see Appendix A.5.

Evaluation. We evaluate performance across multiple multimodal reasoning benchmarks, including MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2023), WeMath (Qiao et al., 2024), HallusionBench (Guan et al., 2024), ChartQA (Masry et al., 2022), LogicVista (Xiao et al., 2024), ChartMuseum (Tang et al., 2026), MMReason (Yao et al., 2025a), and VisuLogic (Xu et al., 2025). We follow the evaluation protocol of Zhu et al. (2025a) and use Qwen2.5-72B-Instruct (Qwen et al., 2025) to extract final answers from model responses and assess their correctness against reference answers.

We compare DUPL against three controlled baselines: (1) the off-the-shelf base models, (2) the strong RLVR baseline GRPO (Shao et al., 2024), and (3) NoisyRollout (Liu et al., 2025b), a passive data augmentation method. We train NoisyRollout under identical experimental settings as DUPL. For broader context, we also include evaluation results from several external models: R1-Onevision-7B (Yang et al., 2025), OpenVLThinker-7B (Deng et al., 2025), VLAA-Thinker-7B (Chen et al., 2025a), MM-Eureka-Qwen-7B (Meng et al., 2025), and PAPO-7B (Wang et al., 2025b).

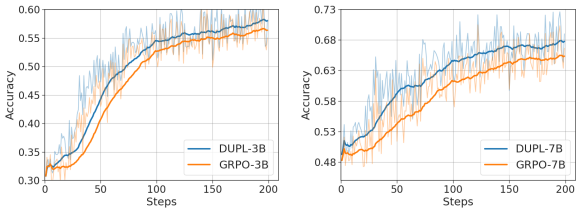
3.2 Main Results

We first evaluate DUPL across six benchmarks encompassing mathematical (MathVerse, MathVista, WeMath) and general-domain tasks (HallusionBench, ChartQA, LogicVista), as shown in Table 1. Compared to the base models, DUPL improves accuracy by up to 12.3% (Avg. 8.7%) for the 3B model and up to 7.9% (Avg. 5.8%) for the 7B model across all evaluated tasks. Furthermore, DUPL consistently outperforms GRPO and NoisyRollout, with DUPL-7B achieving the best overall average performance. Consistent with these results, the training accuracy curves in Figure 2 show that DUPL maintains higher rewards than GRPO throughout training for both model scales, suggesting more effective policy learning.

We also evaluate DUPL on three additional

Model	MathVerse	MathVista	WeMath	HalluBench	ChartQA	LogicVista	Avg.
Open Source Models							
R1-Onevision-7B (Yang et al., 2025)	46.0	63.9	61.8	67.2	78.3	45.5	60.5
OpenVLThinker-7B (Deng et al., 2025)	48.0	70.0	67.1	60.0	78.9	47.1	61.9
VLAA-Thinker-7B (Chen et al., 2025a)	48.2	68.3	67.7	70.0	80.2	47.3	63.6
MM-Eureka-Qwen-7B (Meng et al., 2025)	50.3	71.2	65.6	66.4	79.9	47.3	63.5
PAPO-7B (Wang et al., 2025b)	52.0	73.7	67.6	71.0	79.6	47.1	65.2
DUPL							
Qwen2.5-VL-3B (Bai et al., 2025b)	34.0	58.4	51.8	59.9	73.1	38.0	52.5
+ GRPO	40.6	66.4	60.8	65.5	77.6	39.3	58.4
+ NoisyRollout	41.8	67.7	62.0	66.1	77.7	42.3	59.6
+ DUPL (Ours)	**43.3**	**69.4**	**64.1**	**67.0**	**78.4**	**45.0**	**61.2**
Qwen2.5-VL-7B (Bai et al., 2025b)	45.8	67.2	63.2	65.2	79.8	45.5	61.1
+ GRPO	48.0	70.6	68.5	68.6	81.5	46.0	63.9
+ NoisyRollout	51.0	72.6	69.6	70.1	81.8	47.5	65.4
+ DUPL (Ours)	**52.1**	**74.2**	**71.1**	**71.0**	**84.0**	**48.7**	**66.9**

Table 1: **Model accuracy on diverse visual mathematical and general-domain reasoning benchmarks.** Compared to the base models, DUPL improves accuracy by up to 12.3% (Avg. 8.7%) for the 3B model and up to 7.9% (Avg. 5.8%) for the 7B model across all evaluated tasks. Furthermore, DUPL consistently outperforms the strong baselines GRPO and NoisyRollout, with DUPL-7B achieving the best overall average performance.



(a) 3B models (b) 7B models

Figure 2: **Training accuracy of DUPL and GRPO on 3B and 7B models**. DUPL consistently achieves higher rewards than GRPO throughout training.

benchmarks: ChartMuseum (Tang et al., 2026), MMReason (Yao et al., 2025a), and VisuLogic (Xu et al., 2025). As detailed in Table 14 (Appendix A.9), DUPL improves the 7B base model by an average of 4.0% across these benchmarks and outperforms the GRPO baseline. We report the training cost in Table 13 (Appendix A.7), DUPL incurs only a modest computation overhead relative to GRPO, this minor trade-off is well justified by its large performance gains over the base models and its consistent outperformance of GRPO. Furthermore, we evaluate DUPL across six benchmarks using four random seeds (Table 15, Appendix A.10). Results show that DUPL consistently enhances the 7B model and outperforms GRPO, showing its statistical significance.

Additionally, we evaluate our framework on Qwen3-VL-Instruct 4B and 8B models. As shown in Table 16 (Appendix A.11), DUPL improves accuracy by up to 10.7% (Avg. 5.3%) for the 4B base model and up to 12.4% (Avg. 6.3%) for the 8B base model across all evaluated tasks, while consistently outperforming GRPO.

Taken together, these results demonstrate that DUPL yields consistent and robust improvements across both mathematical and general-domain multimodal reasoning tasks. The observed gains demonstrate that guiding exploration with perceptual and output uncertainty enables more effective policy learning.

3.3 Ablation Studies

In this section, we conduct ablation studies to validate the contribution of each component in DUPL, including perceptual uncertainty, output uncertainty, and the dynamic branch prioritization strategy, as well as the impact of alternative divergence measures. Furthermore, to demonstrate the generalizability of our approach, we extend the experiments: we evaluate DUPL using an alternative RL algorithm, DAPO (Yu et al., 2026), and apply it to a different base model, LLaVA-OneVision-1.5-8B-Instruct (An et al., 2025).

Perceptual Uncertainty. We first investigate the specific contribution of perceptual uncertainty by performing an ablation study where the term u_{per} is deactivated. We report the quantitative results for mathematical reasoning and general-domain benchmarks in Table 2 and Table 3, respectively. As shown, removing perceptual uncertainty from the feedback guidance results in an accuracy drop of up to 3.8% (Avg. 1.7%) on mathematical reasoning benchmarks. On general-domain reasoning tasks, performance decreases by up to 1.3% (Avg. 1.0%). We further illustrate the training dynamics in Figure 5 (Appendix A.8). Without

Approach	MathVerse	MathVista	WeMath	Avg.
Full approach	52.1	74.2	71.1	65.8
w/o u_{per}	48.3	73.6	70.3	64.1
w/o u_{out}	48.6	73.5	70.8	64.3
w/o u_{per} & u_{out}	48.0	73.1	68.5	63.2

Table 2: **Ablations of DUPL accuracy on *mathematical* reasoning benchmarks**. Removing perceptual uncertainty u_{per} or output uncertainty u_{out} reduces performance, with the largest drop when both are removed, confirming their complementary benefits.

Approach	HalluBench	ChartQA	LogicVista	Avg.
Full approach	71.0	84.0	48.7	67.9
w/o u_{per}	69.7	83.4	47.8	66.9
w/o u_{out}	70.2	82.4	47.8	66.8
w/o u_{per} & u_{out}	69.2	82.1	46.4	65.9

Table 3: **Ablations of DUPL on *general-domain* reasoning benchmarks**. The same ablation trend in mathematical reasoning holds in general-domain settings, indicating that perceptual uncertainty and output uncertainty generalize beyond math tasks.

perceptual uncertainty, the learning curve consistently lags behind that of the full DUPL approach, indicating slower and less effective policy optimization. These results confirm that incorporating perceptual uncertainty as a feedback signal plays a critical role in guiding policy learning. By explicitly steering exploration toward visually uncertain states, perceptual uncertainty enables more targeted exploration, thereby improving the model's reasoning accuracy.

Output Uncertainty. Next, we evaluate the impact of output uncertainty by removing the term u_{out}. We report the results on visual mathematical reasoning benchmarks in Table 2, and the results on general-domain reasoning in Table 3. In the absence of output uncertainty guidance, accuracy drops by up to 3.5% (Avg. 1.5%) on mathematical benchmarks and by up to 1.6% (Avg. 1.1%) on general-domain tasks. As illustrated by the training accuracy curves in Figure 5 (Appendix A.8), the variant without output uncertainty lags behind compared to the full DUPL approach.

Furthermore, jointly removing both perceptual and output uncertainty results in a more pronounced performance degradation. As shown in Tables 2 and 3, this combined ablation leads to accuracy drops of up to 4.1% (Avg. 2.6%) on mathematical reasoning benchmarks and up to 2.3% (Avg. 2.0%) on general-domain reasoning tasks.

These results demonstrate that both uncertainties provide effective and complementary feedback signals for policy optimization. Perceptual uncertainty encourages targeted exploration within the visual state space, while output uncertainty induces beneficial stochasticity in the textual output space.

Approach	MathVerse	MathVista	WeMath	Avg.
Forward KL	39.4	70.7	56.1	55.4
Symmetric KL	52.1	74.2	71.1	65.8

Table 4: **DUPL accuracy on *mathematical* reasoning benchmarks under different divergence measures.** Symmetric KL provides stable uncertainty guidance and improves accuracy, while forward KL induces excessive divergence and degrades performance.

Approach	HalluBench	ChartQA	LogicVista	Avg.
Forward KL	67.5	80.3	45.1	64.3
Symmetric KL	71.0	84.0	48.7	67.9

Table 5: **DUPL accuracy on *general-domain* reasoning benchmarks under different divergence measures.** Symmetric KL enables stable uncertainty-guided learning, while forward KL leads to excessive divergence and degraded performance.

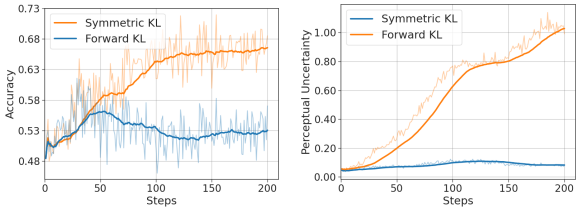
Alternative Divergence Measures. To ablate the formulation of perceptual uncertainty, we replace the symmetric KL divergence with a forward KL variant. However, as shown in Figure 3a, the forward KL divergence leads to unstable training with accuracy declining. We qualitatively analyze the evolution of perceptual uncertainty under different KL formulations throughout training. As illustrated in Figure 3b, perceptual uncertainty measured by forward KL grows excessively large, whereas uncertainty computed using symmetric KL exhibits a moderate increase followed by a gradual decrease, remaining overall stable during training. This observation explains the degraded performance under forward KL: it encourages the model to diverge excessively, leading to unstable policy updates. We also evaluate both variants on diverse reasoning benchmarks (Tables 4 and 5). Consistent with Figure 3, the forward KL formulation results in lower accuracy.

Overall, this ablation validates the use of symmetric KL for measuring perceptual uncertainty, as it provides a stable uncertainty signal that supports effective exploration while preserving training stability.

Dynamic Branch Prioritization. We evaluate the effect of the dynamic branch prioritization mechanism, which gradually adjusts the probability of selecting the noisy branch versus the raw branch. To isolate its effect, we replace it with a fixed sampling probability of 0.5. We report the

Model	MathVerse	MathVista	WeMath	HalluBench	ChartQA	LogicVista	Avg.
Qwen2.5-VL-7B	45.8	67.2	63.2	65.2	79.8	45.5	61.1
+ DAPO	50.0	74.1	69.5	69.8	82.2	48.9	65.8
+ DUPL (w/ DAPO)	**52.8**	**75.7**	**69.6**	**71.2**	**84.1**	**52.0**	**67.6**

Table 6: **Generalization to alternative RL algorithms.** When applied to the DAPO framework, DUPL improves the 7B base model by an average of 6.5% across six benchmarks, consistently outperforming the DAPO baseline.



(a) Training accuracy (b) Perceptual uncertainty

Figure 3: **Training accuracy and qualitative analysis of perceptual uncertainty under different divergence measures.** Measuring perceptual uncertainty with symmetric KL provides a stable signal that effectively guides policy learning while maintaining training stability. In contrast, forward KL produces excessively large perceptual uncertainty, resulting in unstable training and decreased accuracy.

results on mathematical and general-domain reasoning benchmarks in Table 7 and Table 8, respectively. As shown, dynamic branch prioritization consistently outperforms fixed-probability sampling, yielding accuracy improvements of up to 3.6% (Avg. 2.5%) on mathematical reasoning benchmarks and up to 1.8% (Avg. 1.4%) on general-domain reasoning benchmarks. Similar trends are observed in the training curves shown in Figure 6 (Appendix A.8).

These results highlight the advantage of dynamic branch prioritization: higher reliance on the noisy branch facilitates exploration during early training, while progressively emphasizing the raw branch stabilizes optimization and improves convergence in later stages.

Approach	MathVerse	MathVista	WeMath	Avg.
Fixed Prob	48.5	73.6	67.8	63.3
Branch Prioritization	52.1	74.2	71.1	65.8

Table 7: **DUPL accuracy on *mathematical* reasoning benchmarks evaluating dynamic branch prioritization.** Dynamic branch prioritization outperforms fixed-probability sampling.

Alternative RL Algorithms. To demonstrate that DUPL generalizes beyond GRPO, we evaluate it using DAPO as the underlying RL algorithm. As shown in Table 6, integrating DUPL with DAPO improves the 7B base model by an average of 6.5% across six benchmarks, consistently outperforming the standard DAPO baseline. These results highlight that our uncertainty-guided exploration is algorithm-agnostic and can effectively enhance other RLVR methods.

Approach	HalluBench	ChartQA	LogicVista	Avg.
Fixed Prob	69.9	82.6	46.9	66.5
Branch Prioritization	71.0	84.0	48.7	67.9

Table 8: **DUPL accuracy on *general-domain* reasoning benchmarks evaluating dynamic branch prioritization.** Dynamic branch prioritization outperforms fixed-probability sampling.

Alternative Base Models. To evaluate the generalizability of DUPL, we evaluate it on an alternative base model, LLaVA-OneVision-1.5-8B-Instruct (An et al., 2025). As shown in Table 9, our method remains effective, improving the base model by an average of 4.7% across the six benchmarks. Furthermore, DUPL consistently outperforms the GRPO baseline, verifying its broad applicability across diverse architectures.

3.4 Sensitivity Analysis

We conduct a sensitivity analysis of DUPL under varying levels of noise by adjusting the standard deviation of the Gaussian perturbation in the noisy branch ($\sigma \in \{0.2, 0.4, 0.8\}$). As reported in Table 10, moderate noise ($\sigma = 0.4$) achieves the highest accuracy. Low noise ($\sigma = 0.2$) provides insufficient visual exploration, whereas high noise ($\sigma = 0.8$) introduces excessive variance that overly amplifies perceptual uncertainty. This trend is further supported by the training accuracy curves in Figure 4a, as well as the qualitative analysis of perceptual uncertainty evolution over training under different noise levels in Figure 4b.

4 Related Work

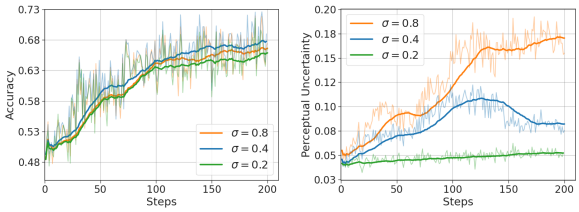
Exploration in Text-Based Reasoning. Recent text-only RLVR works have begun addressing policy exploration. To complement outcome rewards, i-MENTOR (Gao et al., 2025) uses trajectory-aware intrinsic signals and dynamic reward scaling, while Retrospective Replay (Dou et al., 2025) revisits promising early states to counter exploration decay. Other exploration strategies leverage outcome-based schemes (Song et al., 2025)

Model	MathVerse	MathVista	WeMath	HalluBench	ChartQA	LogicVista	Avg.
LLaVA-OneVision-1.5-8B	45.6	68.8	61.6	65.7	81.6	46.0	61.6
+ GRPO	48.0	70.9	67.8	68.5	82.7	46.5	64.1
+ DUPL	**49.8**	**73.5**	**70.6**	**70.8**	**84.8**	**48.2**	**66.3**

Table 9: **Evaluation performance using LLaVA-OneVision-1.5-8B.** DUPL remains effective with this architecture, improving base model accuracy by an average of 4.7% and consistently outperforming GRPO across all tasks.

Approach	MathVerse	MathVista	WeMath	HalluBench	ChartQA	LogicVista	Avg.
$\sigma = 0.2$	48.4	74.0	68.8	69.4	81.9	45.3	64.6
$\sigma = 0.4$	**52.1**	**74.2**	**71.1**	**71.0**	**84.0**	**48.7**	**66.9**
$\sigma = 0.8$	49.2	73.5	66.8	70.4	82.9	46.2	64.8

Table 10: **DUPL accuracy on diverse reasoning benchmarks with different noise levels.** Moderate noise ($\sigma = 0.4$) yields the best accuracy across all evaluated benchmarks.



(a) Training accuracy (b) Perceptual uncertainty

Figure 4: **Training accuracy and perceptual uncertainty across noise levels**. Moderate noise ($\sigma = 0.4$) yields the best performance. Low noise ($\sigma = 0.2$) provides insufficient visual exploration, whereas high noise ($\sigma = 0.8$) introduces excessive variance that overly amplifies perceptual uncertainty.

or granular process rewards (Setlur et al., 2024), though scoring intermediate steps remains challenging. Additionally, upweighting negative samples can mitigate diversity collapse (Zhu et al., 2025b), and EVOL-RL promotes reasoning diversity through novelty-oriented reinforcement without explicit labels (Zhou et al., 2025).

Multimodal Reasoning. RL is increasingly applied to enhance multimodal reasoning. Yang et al. (2025) and Huang et al. (2025) extended language reasoning with visual inputs and vision-grounded prompts, while others focused on cross-modal generalization (Deng et al., 2025) and unifying visual-textual signals (Chen et al., 2025a). Further advances include self-rewarding mechanisms that decompose visual and linguistic reasoning (Li et al., 2025b), perception-aware losses for better visual grounding (Wang et al., 2025b), single-rollout RL for improved training efficiency (Liu et al., 2025a), generative multimodal reward modeling (Liu et al., 2026c), and enhancing reasoning robustness (Liu et al., 2026d).

Despite these advances, effective exploration in multimodal RLVR remains largely underexplored. Existing methods generally treat visual inputs as deterministic, overlooking inherent perceptual ambiguity. While recent works inject noise into training images to encourage exploration (Liu et al., 2025b; Yao et al., 2025b), they rely on passive data augmentation. The policy learning objective remains unchanged, exploration is global and undirected.

In contrast, DUPL introduces targeted exploration for multimodal RLVR via dual-uncertainty guided policy learning. By jointly modeling perceptual uncertainty and output uncertainty and incorporating them into an uncertainty-driven feedback loop, DUPL actively regulates policy updates and directs exploration toward states where the model is uncertain, addressing a key limitation of prior multimodal RLVR methods.

5 Conclusions

In conclusion, we introduce DUPL, a dual-uncertainty guided policy learning approach that overcomes the limitations of undirected exploration in multimodal RLVR. By measuring both perceptual and output uncertainty, DUPL creates a targeted feedback loop, utilizing dynamic branch prioritization and advantage shaping to guide policy updates. Extensive evaluation across diverse mathematical and general-domain benchmarks demonstrate that DUPL yields solid improvements. It achieves accuracy gains of up to 12.3% (3B) and 7.9% (7B) on Qwen2.5-VL, alongside up to 10.7% (4B) and 12.4% (8B) on Qwen3-VL-Instruct. Validating its broad applicability, we further demonstrate that DUPL successfully enhances an alternative base model, LLaVA-OneVision-1.5-8B, by 4.7% on average, and effectively generalizes to the DAPO algorithm with an average gain of 6.5%. These results establish DUPL as an effective approach of targeted explo-

ration for multimodal RLVR.

Limitations

While DUPL effectively demonstrates the value of dual-uncertainty guided policy learning, it presents several promising avenues for future exploration. Consistent with standard practice in prior work (Yarats et al., 2021; Laskin et al., 2020; Liu et al., 2025b), our current implementation employs augmentation techniques such as geometric manipulations and Gaussian noise, to induce perceptual uncertainty. A promising avenue for future investigation lies in exploring more sophisticated perturbation strategies or learnable noise generators that could expose subtler reasoning vulnerabilities. Furthermore, extending the dual-uncertainty framework to dynamic modalities, such as video, presents a worthy direction for establishing a more comprehensive future work.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. 2025b. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.

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*.

Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, and Dong Yu. 2025. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *Preprint*, arXiv:2509.09675.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *Preprint*, arXiv:2503.17352.

Shihan Dou, Muling Wu, Jingwen Xu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Improving rl exploration for llm reasoning through retrospective replay. *arXiv preprint arXiv:2504.14363*.

Jingtong Gao, Ling Pan, Yejing Wang, Rui Zhong, Chi Lu, Qingpeng Cai, Peng Jiang, and Xiangyu Zhao. 2025. Navigate the unknown: Enhancing llm reasoning with intrinsic motivation guided exploration. *arXiv preprint arXiv:2505.17621*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

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

Håkan Hjalmarsson. 2005. From experiment design to closed-loop control. *Automatica*, 41(3):393–438.

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

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. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. 2020. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. 2025a. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and 1 others. 2025b. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*.

Rui Liu, Peng Gao, Yu Shen, Ming Lin, and Pratap Tokekar. 2026a. Adaptive conformal guidance for learning under uncertainty. In *International Conference on Learning Representations*, volume 2026, pages 6168–6192.

Rui Liu, Guangyao Shi, and Pratap Tokekar. 2023. Data-driven distributionally robust optimal control with state-dependent noise. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9986–9991. IEEE.

Rui Liu, Pratap Tokekar, and Ming Lin. 2026b. Active asymmetric multi-agent multimodal learning under uncertainty. *arXiv preprint arXiv:2602.04763*.

Rui Liu, Dian Yu, Lei Ke, Haolin Liu, Yujun Zhou, Zhenwen Liang, Haitao Mi, Pratap Tokekar, and Dong Yu. 2025a. Stable and efficient single-rollout rl for multimodal reasoning. *arXiv preprint arXiv:2512.18215*.

Rui Liu, Dian Yu, Zhenwen Liang, Yucheng Shi, Tong Zheng, Runpeng Dai, Haitao Mi, Pratap Tokekar, and 1 others. 2026c. Deltarubric: Generative multimodal reward modeling via joint planning and verification. *arXiv preprint arXiv:2605.09269*.

Rui Liu, Dian Yu, Haolin Liu, Yucheng Shi, Tong Zheng, Runpeng Dai, Haitao Mi, Pratap Tokekar, and 1 others. 2026d. Reinforcing multimodal reasoning against visual degradation. *arXiv preprint arXiv:2605.09262*.

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. *arXiv preprint arXiv:2504.13055*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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 mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

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

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*.

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*.

Yuda Song, Julia Kempe, and Remi Munos. 2025. Outcome-based exploration for llm reasoning. *arXiv preprint arXiv:2509.06941*.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Liu, Zayne Sprague, and 1 others. 2026. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *Advances in Neural Information Processing Systems*, 38.

Christian Walder and Deep Karkhanis. 2025. Pass@k policy optimization: Solving harder reinforcement learning problems. *arXiv preprint arXiv:2505.15201*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

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*.

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

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, and 1 others. 2025a. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward agi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 273–283.

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

Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2021. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International conference on learning representations*.

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.

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. 2025. Entropy-based exploration conduction for multi-step reasoning. *arXiv preprint arXiv:2503.15848*.

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.

Tong Zheng, Hongming Zhang, Wenhao Yu, Xiaoyang Wang, Xinyu Yang, Runpeng Dai, Rui Liu, Huiwen Bao, Chengsong Huang, Heng Huang, and 1 others. 2025a. Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*.

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

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*.

Linghao Zhu, Yiran Guan, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Bin Qin, Jian Luan, Yuliang Liu, and Xiang Bai. 2025a. Shuffle-r1: Efficient rl framework for multimodal large language models via data-centric dynamic shuffle. *arXiv preprint arXiv:2508.05612*.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025b. The surprising effectiveness of negative reinforcement in llm reasoning. *arXiv preprint arXiv:2506.01347*.

A Appendix

A.1 GRPO Preliminaries

In GRPO, given an input x, a group of responses $\{o_i\}_{i=1}^G$ are sampled from the old policy $\pi_{\theta_{\text{old}}}$, each associated with a reward r_i. Then the normalized advantage for response o_i is defined as:

$$A_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^G)}{\text{std}(\{r_i\}_{i=1}^G)}. \tag{2}$$

As in PPO, GRPO uses clipped importance sampling to stabilize policy updates. Let $\rho_i(\theta) = \frac{\pi_\theta(o_i|x)}{\pi_{\theta_{\text{old}}}(o_i|x)}$ denote the probability ratio between the new and old policies. The GRPO objective is to maximize the following equation:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{o_i\}\sim\pi_{\theta_{\text{old}}}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i(\theta)A_i, \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)A_i\Big)\Bigg], \tag{3}$$

where ϵ_{clip} is the clipping hyperparameter.

A.2 Algorithm

We present the full procedure of DUPL in Algorithm 1, summarizing Section 2. First, we measure output and perceptual uncertainty, which are then used to compute guidance signals. These signals are employed to shape the advantage for the raw and noisy branches separately. A dynamic branch prioritization mechanism is then applied to select which uncertainty-guided advantage is used for policy updates at each training step.

A.3 Analysis

We express the policy gradient for the noisy branch as follows:

$$\nabla_\theta \mathcal{J}_{\text{DUPL}}(\theta) \propto \mathbb{E}_{o'\sim\pi_{\theta_{\text{old}}}}\Big[(A^{\text{noi}} + g_{\text{out}}^{\text{noi}} + g_{\text{per}}) \cdot \nabla_\theta \log \pi_\theta(o' \mid x')\Big]. \tag{4}$$

From a gradient perspective, this formulation encourages more effective policy updates for targeted exploration compared to standard RLVR approaches such as GRPO. We omit caps/clipping for clarity. The term g_{per} explicitly reweights the gradient to favor trajectories originating from states with high perceptual ambiguity, guiding the model to acquire more informative visual features. The term $g_{\text{out}}^{\text{noi}}$ acts as a general-purpose exploration mechanism in the token space: by rewarding higher entropy in the output distribution, it prevents premature convergence and complements the visual exploration induced by g_{per}. Unlike GRPO, which rely solely on the extrinsic reward signal A^{noi}, this decomposition shows that DUPL optimizes a composite objective that enables targeted exploration.

A.4 Prompt Templates

We list below the prompt used to instruct the model to produce the structured outputs.

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{}`.

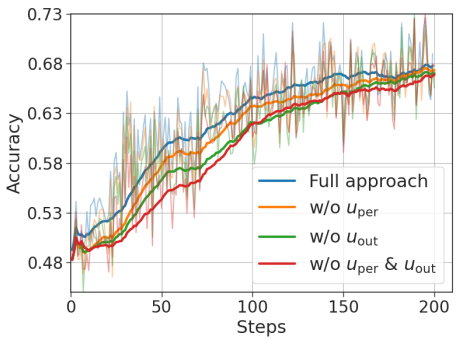


Figure 5: **Training dynamics of accuracy rewards under perceptual uncertainty u_{per} and output uncertainty u_{out}.** Each uncertainty signal provides effective feedback guidance and improves performance, while their combination of full approach yields the best results.

A.5 Training Details

We train all models on the MMRL30k dataset (Zhu et al., 2025a), which contains around 30K samples. The models are trained to generate responses in a structured format, where the reasoning process is enclosed within `<think></think>` tags and the final answer is presented in `\boxed{}`. The reward function for RL training is a combination of a format reward and an accuracy reward, with coefficients

Algorithm 1 Dual-Uncertainty Guided Policy Learning (DUPL)

Require: Dataset $\mathcal{D}$, group size G, total training steps s_{total}, augmentation function $\mathcal{T}$.

Initialize policy parameters θ, old policy $\theta_{\text{old}} \leftarrow \theta$
for $s = 1$ to s_{total} **do**
  Sample input $x = (x_{\text{text}}, x_{\text{image}}) \sim \mathcal{D}$
  Construct perturbed image $x'_{\text{image}} = \mathcal{T}(x_{\text{image}})$, $x' = (x_{\text{text}}, x'_{\text{image}})$
  Sample group of responses $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$, $\{o'_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x')$
  Compute branch prioritization probability: $p_{\text{noi}}(s) = \max\left(0, 1 - \frac{s}{s_{\text{total}}}\right)$
  for $i = 1$ to G **do**
    Compute raw advantage A_i^{raw}, noisy advantage A_i^{noi} (Eq. 2)
    Compute output uncertainty $u_{\text{out},i}^{\text{raw}}$ and $u_{\text{out},i}^{\text{noi}}$ for each branch
    Compute perceptual uncertainty $u_{\text{per},i}$ using token-level symmetric KL divergence (Eq. 1)
    Compute guidance signals:

$$g_{\text{per},i} = \min\left(\frac{|A_i^{\text{noi}}|}{\beta_p}, \alpha_p \cdot \text{stopgrad}(u_{\text{per},i})\right)$$

$$g_{\text{out},i}^{\text{raw}} = \min\left(\frac{|A_i^{\text{raw}}|}{\beta_o}, \alpha_o \cdot \text{stopgrad}(u_{\text{out},i}^{\text{raw}})\right), g_{\text{out},i}^{\text{noi}} = \min\left(\frac{|A_i^{\text{noi}}|}{\beta_o}, \alpha_o \cdot \text{stopgrad}(u_{\text{out},i}^{\text{noi}})\right)$$

    Compute shaped advantages:

$$\widehat{A}_i^{\text{raw}} = A_i^{\text{raw}} + g_{\text{out},i}^{\text{raw}}, \quad \widehat{A}_i^{\text{noi}} = A_i^{\text{noi}} + g_{\text{out},i}^{\text{noi}} + g_{\text{per},i}$$

    Sample branch selector $z_i \sim \text{Bernoulli}(p_{\text{noi}}(s))$
    Select final advantage (with corresponding selected (x, o_i) or (x', o'_i)):

$$\widehat{A}_i = z_i \cdot \widehat{A}_i^{\text{noi}} + (1 - z_i) \cdot \widehat{A}_i^{\text{raw}}$$

  end for
  Compute surrogate objective with shaped advantages (Eq. 3)
  Update policy parameters $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}_{\text{DUPL}}(\theta)$
end for

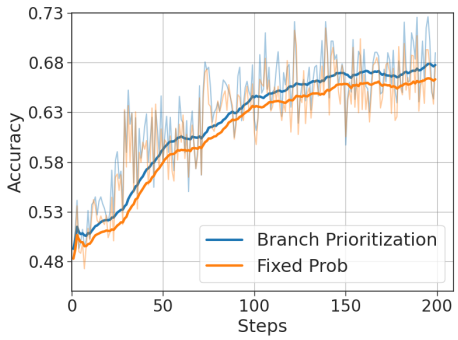


Figure 6: **Training dynamics of accuracy rewards comparing dynamic branch prioritization with fixed-probability sampling.** Dynamic branch prioritization consistently achieves higher accuracy.

of 0.1 and 0.9, respectively. The training is performed for 200 steps using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $1e-6$ and a weight decay of 0.01. We adopt a global batch size of 128, a rollout batch size of 256, and generate 5 rollouts per input with a rollout temperature 1.0. The implementation builds on the framework EasyR1 (Zheng et al., 2025b).

When transferring perceptual and output uncertainty into guidance signals for advantage shaping to guide policy learning, we use $\alpha_p = 1.0$ and $\beta_p = 2.0$ for perceptual uncertainty, and $\alpha_o = 0.4$ and $\beta_o = 2.0$ for output uncertainty. A detailed sensitivity analysis of these parameters is provided in Appendix A.6.

A.6 Additional Sensitivity Analysis

To investigate the impact of the uncertainty feedback guidance parameters on advantage shaping, we conduct a systematic sensitivity analysis via grid search using the Qwen2.5-VL-7B model. Our framework introduces two sets of scaling factors: (α_o, β_o), which regulate the token-level output en-

tropy guidance g_{out}, and (α_p, β_p), which modulate the symmetric KL divergence guidance g_{per} for capturing perceptual uncertainty.

We first analyze the effect of the output uncertainty parameters by varying the scaling factor $\alpha_o \in \{0.2, 0.4, 0.8\}$ and $\beta_o \in \{1.0, 2.0, 4.0\}$. As shown in Table 11, moderate values ($\alpha_o = 0.4, \beta_o = 2.0$) yield the best performance. Setting α_o low restricts action-space stochasticity, whereas high values introduce noise into the advantage estimation, both of which degrade final reasoning accuracy. Next, we analyze the performance of DUPL under varying perceptual uncertainty parameters by sweeping the coefficients $\alpha_p \in \{0.5, 1.0, 2.0\}$ and $\beta_p \in \{1.0, 2.0, 4.0\}$. As demonstrated in Table 12, the configuration of $\alpha_p = 1.0$ and $\beta_p = 2.0$ achieves the highest accuracy.

α_o	β_o	MathVerse	MathVista	ChartQA	Avg.
0.2	2.0	50.1	72.4	82.1	68.2
0.4	**2.0**	**52.1**	**74.2**	**84.0**	**70.1**
0.8	2.0	48.9	71.5	81.3	67.2
0.4	1.0	51.8	73.0	83.1	69.3
0.4	4.0	51.2	73.5	83.4	69.4

Table 11: **Sensitivity analysis of output uncertainty parameters** (α_o, β_o). $\alpha_o = 0.4$ and $\beta_o = 2.0$ yield the best performance.

α_p	β_p	MathVerse	MathVista	ChartQA	Avg.
0.5	2.0	50.5	72.9	81.8	68.4
1.0	**2.0**	**52.1**	**74.2**	**84.0**	**70.1**
2.0	2.0	50.4	72.2	81.9	68.2
1.0	1.0	51.0	73.2	82.9	69.0
1.0	4.0	51.5	73.6	83.5	69.5

Table 12: **Sensitivity analysis of perceptual uncertainty parameters** (α_p, β_p). $\alpha_p = 1.0$ and $\beta_p = 2.0$ yield the best performance.

A.7 Training Cost Analysis

We compare the training cost of DUPL and GRPO by measuring training time (minutes per step) and throughput (tokens processed per second per GPU). As shown in Table 13, DUPL incurs only a modest computational overhead compared to GRPO, reflected in a slight increase in training time per step and a corresponding reduction in throughput. However, this minor overhead is well justified by the large performance gains DUPL provides over the base models and its consistent outperformance of the GRPO baseline.

Method	Time (mins/step)	Throughput (tokens/s/GPU)
GRPO	4.12	496.14
DUPL	4.95	453.70

Table 13: **Training cost comparison.** DUPL incurs only a modest computational overhead relative to GRPO, resulting in slightly increased training time per step and a corresponding reduction in throughput. This minor trade-off is well justified by the large performance gains DUPL achieves over the base models and its consistent outperformance of the GRPO baseline.

A.8 Ablation Studies

In Section 3.3, we conduct a comprehensive set of ablation studies to validate the contribution of each component in DUPL, including perceptual uncertainty, output uncertainty, the dynamic branch prioritization strategy, and the choice of divergence measures. Here, we present additional results that further validate the role of some components.

In Figure 5, we show the training dynamics of accuracy under perceptual uncertainty u_{per} and output uncertainty u_{out}. Each uncertainty signal provides effective feedback guidance and improves performance, while their combination in the full approach yields the best results.

We present the training dynamics of accuracy in Figure 6, comparing dynamic branch prioritization with fixed-probability sampling. We observe that dynamic branch prioritization consistently achieves higher accuracy throughout training. These results highlight its advantage in balancing exploration and stability: increased reliance on the noisy branch facilitates exploration during early training, while progressively emphasizing the raw branch stabilizes optimization and improves convergence in later stages, leading to superior final performance.

A.9 Evaluation on More Benchmarks

We evaluate on additional benchmarks, including ChartMuseum (Tang et al., 2026), MMReason (Yao et al., 2025a), and VisuLogic (Xu et al., 2025). As shown, DUPL improves the base 7B model by average 4.0% and outperforms GRPO.

A.10 Statistical Significance

We train the models across four random seeds and evaluate on six benchmarks, reporting the mean and standard deviation in Table 15. DUPL consistently outperforms the 7B base model and the

Model	ChartMuseum	MMReason	VisuLogic	Avg.
Qwen2.5-VL-7B	26.8	16.8	26.0	23.2
+ GRPO	28.8	18.9	28.5	25.4
+ DUPL	**31.3**	**20.5**	**29.8**	**27.2**

Table 14: **Evaluation on additional multimodal reasoning benchmarks.** DUPL improves the 7B base model by an average of 4.0% and outperforms the GRPO baseline.

GRPO baseline, confirming both statistical significance and robust training stability.

Benchmark	Base 7B	GRPO	DUPL (Ours)
MathVerse	45.8	48.1 $\pm$ 0.2	**52.2** $\pm$ 0.3
MathVista	67.2	70.5 $\pm$ 0.4	**74.3** $\pm$ 0.2
WeMath	63.2	68.4 $\pm$ 0.1	**71.1** $\pm$ 0.2
HalluBench	65.2	68.7 $\pm$ 0.3	**71.2** $\pm$ 0.3
ChartQA	79.8	81.8 $\pm$ 0.2	**83.9** $\pm$ 0.1
LogicVista	45.5	46.2 $\pm$ 0.3	**48.8** $\pm$ 0.2
Average	61.1	64.0	**66.9**

Table 15: **Evaluation on six benchmarks across four random seeds**. We report mean and standard deviation. DUPL consistently improves over GRPO and the base model, demonstrating both performance gains and statistical significance.

A.11 Evaluation with Qwen3-VL

We evaluate our framework using Qwen3-VL-Instruct 4B and 8B models. As shown in Table 16, DUPL improves accuracy by up to 10.7% (Avg. 5.3%) for the 4B base model and up to 12.4% (Avg. 6.3%) for the 8B base model across all evaluated tasks. Furthermore, DUPL consistently outperforms the strong GRPO baseline, with DUPL-8B achieving the best overall average performance.

Model	MathVerse	MathVista	WeMath	HalluBench	ChartQA	Avg.
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	57.6	75.5	65.9	73.4	79.5	70.4
+ GRPO	62.0	78.2	76.3	74.7	81.7	74.6
+ DUPL (Ours)	**65.1**	**78.4**	**76.6**	**75.7**	**82.8**	**75.7**
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	58.6	76.8	70.5	74.4	79.6	72.0
+ GRPO	63.4	78.9	77.6	75.6	82.2	75.5
+ DUPL (Ours)	**68.1**	**80.0**	**82.9**	**76.6**	**84.1**	**78.3**

Table 16: **Evaluation performance using Qwen3-VL models.** Compared to the base models, DUPL improves accuracy by up to 10.7% (Avg. 5.3%) for the 4B model and up to 12.4% (Avg. 6.3%) for the 8B model across all evaluated tasks. Furthermore, DUPL consistently outperforms the strong GRPO baseline, with DUPL-8B achieving the best overall average performance.