MaPP: A Unified Marginalized Posterior-Predictive Framework for Data-Efficient RLVR

Yangyang Ren[1,2*] **Haodong Zhu**[1,2*] **Sheng Xu**[3†] **Yanjing Li**[4†]
Nikolai Yu. Zolotykh[5] **Wentao Zhang**[2,6] **Baochang Zhang**[1,7]

[1]Beihang University [2]Zhongguancun Academy [3]Communication University of China
[4]Nanyang Technological University [5]Lobachebsky University
[6]Peking University [7]Hangzhou Innovation Institute of Beihang University

Abstract

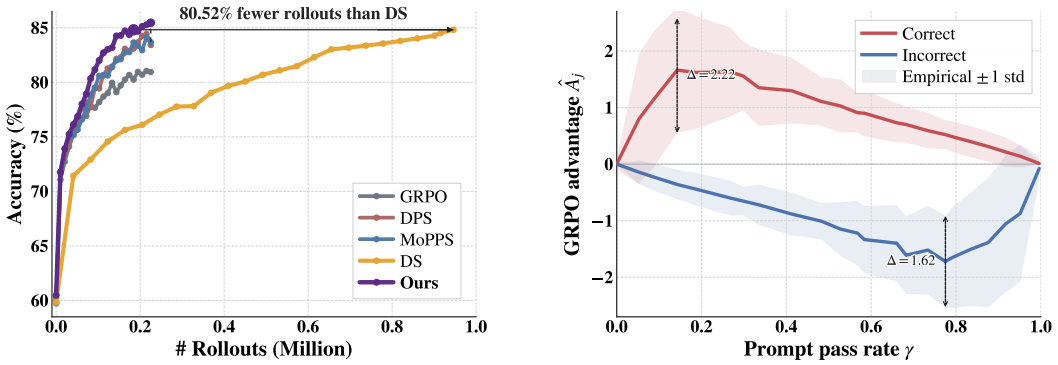
Reinforcement learning with verifiable rewards (RLVR) enhances the reasoning capabilities of large language models (LLMs), but at the expense of significant computational overhead due to compute-intensive rollout processes and frequent policy updates. Online prompt selection, a widely adopted strategy for improving training efficiency, maintains per-prompt Bayesian posteriors to predict prompt difficulty and prioritize informative prompts before committing rollout budget. However, these methods assess prompt informativeness without accounting for how reliably learning signal is extracted from sampled responses. In GRPO-based RL algorithms, the realized advantage of a response depends not only on its own outcome, but also on the randomly sampled outcomes of its peers through group normalization. Our experimental and theoretical analysis show that the resulting uncertainty in group composition induces *composition noise*, a non-vanishing variance component that creates an irreducible lower bound on gradient estimation error. Consequently, the standard group-relative advantage fails to faithfully characterize response-level utility, degrading gradient estimation and thereby impairing downstream prompt selection. To address this issue, we propose a unified ***Ma**rginalized **P**osterior-**P**redictive* framework (***MaPP***) for data-efficient RLVR, which first denoises response-level advantage estimation and then improves prompt selection using a shared Beta posterior. Specifically, for each response, MaPP replaces the standard group-relative advantage with a composition-invariant intrinsic advantage via closed-form Beta–Binomial marginalization. This yields a closed-form posterior-predictive advantage estimator whose error provably diminishes as the posterior concentrates. Then, building on the same posterior, MaPP derives an uncertainty-aware prompt selection score that more faithfully characterizes prompt informativeness, improving data efficiency without additional rollout cost. Experiments across mathematics, planning, and visual geometry on five model backbones show that MaPP consistently outperforms GRPO and strong selection baselines, achieving up to +2.45 average accuracy improvement over the strongest baseline under the same rollout budget, a new state-of-the-art.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) is a key post-training paradigm for enhancing LLM reasoning [Guo et al., 2025; Jaech et al., 2024; Shao et al., 2024; Yang et al., 2025; Yu et al., 2025]. Among RLVR algorithms, GRPO [Shao et al., 2024] is most widely adopted, eliminating a

*Equal contribution.
†Corresponding authors: Sheng Xu (shengxu@cuc.edu.cn) and Yanjing Li (yanjing.li@ntu.edu.sg).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Test Accuracy vs. Rollout Cost

(b) Composition Noise in GRPO Advantages

Figure 1: On the CountDown dataset using the Qwen3-4B model, (a) MaPP achieves higher accuracy than all selection methods with the same rollout cost, while requiring 80% fewer rollouts than DS to reach comparable performance. (b) Group-composition uncertainty induces *composition noise* in GRPO (group size G=8; mid-training checkpoint of MoPPS; 100 prompts $\times$ 50 groups): under the same policy checkpoint, identical correct (red) or incorrect (blue) response receives widely varying advantages depending on group composition ($\pm$1 std shaded). The noise persists across all non-degenerate difficulty levels, *e.g.*, $\Delta = 2.22$ at $\gamma \approx 0.15$ and $\Delta = 1.62$ at $\gamma \approx 0.8$.

learned value network by estimating advantages via within-group normalization of G responses per prompt. However, RLVR incurs high computation and memory costs due to intensive rollouts for policy evaluation and updates [Lin et al., 2025; Zheng et al., 2025c]. Thus, improving utility per fixed rollout budget is a central challenge in scaling RLVR.

Existing methods improve training efficiency by selecting prompts for rollout budget. Uniform sampling is inefficient: overly easy or hard prompts cause degenerate groups (all correct or incorrect), wasting rollouts on zero gradient [Bae et al., 2026; Chen et al., 2025; Zeng et al., 2025]. Only intermediate-difficulty prompts provide informative signals. Dynamic Sampling [Yu et al., 2025] oversamples a candidate set, rolls out all, then discards degenerate ones—but the rollouts on discarded prompts are already spent [Zheng et al., 2025c].

To avoid this waste, prediction-based methods [Qu et al., 2026a; Zheng et al., 2025b; Zhou et al., 2026] estimate prompt difficulty *before* rollout and skip likely degenerate prompts entirely, achieving comparable or superior performance with significantly fewer rollouts, as shown in Fig. 1 (a). However, conventional methods which select prompt before rollout do not by itself determine how reliably learning signal is extracted from the responses sampled afterward. In GRPO-based RLVR [Shao et al., 2024], both prompt selection and policy optimization are driven by group-level statistics formed from the sampled responses. For non-degenerate prompts, these counts are not fixed properties of the prompt itself, but random rewards with uncertainty induced by the current policy's latent success rate under a finite number of sampled responses. As a result, even for the same response on the same prompt under the same policy, the realized advantage can deviate substantially depending on which peers happen to be drawn. Fig. 1 (b) visualizes this effect: the same correct or incorrect response receives widely varying advantages across different sampled groups, despite sharing the same prompt and policy. We refer to this deviation, induced by uncertainty in group composition, as *composition noise*. This composition noise degrades gradient estimation and, in turn, limits how faithfully prompt informativeness can be assessed for downstream selection.

Crucially, the information needed to account for this noise already exists in current selection pipelines: the per-prompt Beta posterior maintained over the latent pass rate γ_τ tracks exactly the quantity that governs group-composition uncertainty. To leverage this observation, we propose a ***Ma**rginalized **P**osterior-**P**redictive* framework (***MaPP***), which first denoises response-level advantage estimation and then improves prompt selection using a shared Beta posterior. The key insight is that the GRPO advantage conditions on a single realized group composition, whereas the *true* learning signal of a response is determined by its correctness and the prompt's intrinsic difficulty alone. The first component of MaPP, which we call **MaPP-AD**, denoises the realized GRPO advantage via closed-form Beta–Binomial marginalization, recovering a composition-invariant *intrinsic advantage*. This yields a closed-form posterior-predictive advantage estimator whose error provably diminishes as the posterior concentrates. Building on the same posterior, the second component of MaPP, which

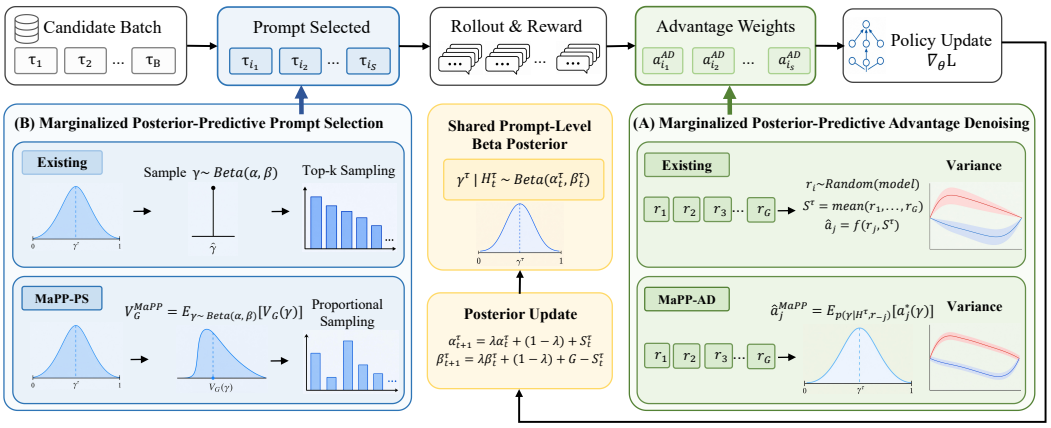


Figure 2: **Overview of the MaPP framework.** A shared Beta posterior $(\alpha_t^\tau, \beta_t^\tau)$, updated online from rollout history, serves as the foundation for two closed-form components. **(A) MaPP-AD** marginalizes $a_j^\star(\gamma)$ over a leave-one-out posterior, replacing the composition-dependent GRPO advantage with composition-invariant weights. Both components reuse the same posterior at no additional rollout cost. **(B) MaPP-PS** marginalizes $V_G(\gamma)$ over the full posterior for uncertainty-aware prompt selection, replacing the point-estimate scoring used by existing methods.

we call **MaPP-PS**, derives a marginalized posterior-predictive, uncertainty-aware prompt selection method that more accurately characterizes prompt informativeness. Both components are modular and can be layered on top of existing selection pipelines. Our contributions are threefold:

- We identify *composition noise*, a within-rollout variance component in the GRPO advantage induced by uncertainty in group composition and show that it degrades gradient estimation and thereby limits downstream prompt selection (§3.2).
- We propose MaPP, a unified posterior-predictive framework with two closed-form components: MaPP-AD, which denoises the realized GRPO advantage via posterior-predictive marginalization, and MaPP-PS, which uses the same shared Beta posterior to derive a more faithful prompt-selection score (§4).
- Experiments across three domains and five model backbones show that MaPP consistently outperforms GRPO and strong selection baselines, achieving up to $+2.45$ average accuracy improvement over the strongest baseline with the same rollout budget.

2 Related Work

2.1 RL Finetuning for Large Language Models

RLVR is a dominant post-training paradigm for reasoning LLMs [Guo et al., 2025; Jaech et al., 2024; Shao et al., 2024]. GRPO [Shao et al., 2024] eliminates PPO's [Schulman et al., 2017] learned value network via within-group normalization, setting the template for many follow-ups. Variants address specific failure modes: Dr.GRPO [Liu et al., 2025] removes length normalization and std scaling, DAPO [Yu et al., 2025] adds asymmetric clipping and dynamic sampling, VAPO [Yue et al., 2025] reintroduces a value model for long CoT. Others include sequence-level importance ratios [Zheng et al., 2025a] and adaptive reweighting [Yang et al., 2026]. These modify *what* is optimized. MaPP is complementary: it modifies *how* the GRPO advantage is estimated under group composition uncertainty, while operating within the same GRPO training framework. Conceptually, MaPP Rao–Blackwellizes [Casella and Robert, 1996] the GRPO advantage using a conjugate Beta posterior instead of a learned critic [Liu et al., 2017].

2.2 Online Prompt Selection for RLVR

Prompts in RLVR contribute unevenly: overly easy or extremely hard prompts yield vanishing gradients, while intermediate ones provide informative signals [Bae et al., 2026; Yu et al., 2025; Zeng et al., 2025]. This motivates online selection methods [Mao et al., 2026; Qu et al., 2026a,b; Yu et al., 2025; Zheng et al., 2025b; Zhou et al., 2026].

Existing approaches fall into two paradigms [Qu et al., 2026b]: evaluation-based (e.g., Dynamic Sampling [Yu et al., 2025], SPEED-RL [Zhang et al., 2025]) oversample and filter, wasting budget; prediction-based (MoPPS [Qu et al., 2026a], DPS [Mao et al., 2026], GRESO [Zheng et al., 2025b], GPS [Qu et al., 2026b], INSIGHT [Zhou et al., 2026]) estimate difficulty before rollout. These mainly improve budget allocation across prompts, but ignore composition noise from uncertainty in group composition when GRPO assigns response-level advantages.

Unlike these methods, MaPP also addresses signal extraction from sampled responses. Specifically, MaPP-AD denoises the GRPO advantage via posterior-predictive marginalization, and MaPP-PS reuses the shared Beta posterior for a more faithful prompt-selection score. Both are modular and can be layered on existing pipelines.

3 Challenge Analysis

3.1 Preliminaries

GRPO for RLVR: Given a prompt τ from training dataset $\mathcal{T} = \{\tau_i\}_{i=1}^{N}$,where N is the size of the training dataset, and G independent responses $\{y_j\}_{j=1}^{G}$ generated by the policy $\pi_\theta(\cdot \mid \tau)$, each receiving a binary reward $r_j^\tau \in \{0, 1\}$ verifying correctness [Guo et al., 2025; Shao et al., 2024], the objective of RL finetuning is to maximize the expected return:

$$\max_\theta \ \mathbb{E}_{\tau\sim\mathcal{T},\ \{y_j\}_{j=1}^{G}\sim\pi_\theta(\cdot|\tau)} \left[\frac{1}{G} \sum_{j=1}^{G} r_j^\tau \right], \tag{1}$$

where G denotes the group size. GRPO [Shao et al., 2024] eliminates the need for a learned value network by estimating advantages through within-group normalization. For each prompt τ, the policy generates a group of G independent responses $\{y_j\}_{j=1}^{G}$, each receiving a binary reward $r_j^\tau \in \{0, 1\}$. The group-relative advantage for the j-th response is then defined as:

$$\hat{A}_j^\tau = \frac{r_j^\tau - \text{mean}(\{r_k^\tau\}_{k=1}^{G})}{\text{std}(\{r_k^\tau\}_{k=1}^{G})}. \tag{2}$$

As noted in DAPO [Yu et al., 2025], when all responses to a given prompt are either all correct or all incorrect, they receive identical rewards, yielding $\hat{A}_j^\tau = 0$. In this scenario, the prompt is *degenerate* and contributes no gradient signal [Yu et al., 2025; Zheng et al., 2025b]. To identify such conditions, prior work on prompt selection [Zheng et al., 2025b] introduces a success count variable $S^\tau \in \{0, G\}$ indicating the degenerate case, *i.e.*, all responses are correct or all are incorrect.

Online Prompt Selection: Prediction-based online prompt selection introduces a success rate $\gamma^\tau \in [0, 1]$ for each prompt τ to estimate prompt difficulty for selection. Since γ is not directly observable, prior works [Qu et al., 2026a; Zhou et al., 2026] first model each response reward $r_j^\tau \in \{0, 1\}$ generated from prompt τ as an independent Bernoulli trial:

$$r_j^\tau \mid \gamma^\tau \sim \text{Bernoulli}(\gamma^\tau) \quad \forall j \in \{1, \cdots G\}. \tag{3}$$

We define the corresponding success count as $S^\tau := \sum_{j=1}^{G} r_j^\tau$, which follows a Binomial distribution, *i.e.*, $S^\tau \mid \gamma^\tau \sim \text{Binomial}(G, \gamma^\tau)$.

Then, a recursive Bayesian update procedure is applied for efficient posterior inference of the success rates γ^τ. At the initial stage a Beta prior models the initial success rate formulated as:

$$\gamma_0^\tau \sim \text{Beta}(\alpha_0^\tau, \beta_0^\tau), \tag{4}$$

where α_0^τ and β_0^τ reflect prior pseudo-counts of successes and failures, typically set to $(1, 1)$ for a uniform prior. During prompt selection, the optimization history is denoted as $\mathcal{H}_t = \{B_t, R_t\}$, where B_t is the selected prompt batch at step t and $R_t = \{r_j^\tau\}_{\tau\in B_t}^{j\in G}$ collects the corresponding rewards. By Bayes rule, the posterior distribution over γ_t^τ given observations up to step t is:

$$p(\gamma_t^\tau \mid \mathcal{H}_t) \propto p(r_j^\tau \mid \gamma_t^\tau) \cdot p(\gamma_t^\tau \mid \mathcal{H}_{t-1}), \tag{5}$$

where $p(\gamma_t^\tau \mid \mathcal{H}_{t-1}) \sim \text{Beta}(\alpha_t^\tau, \beta_t^\tau)$ represents the conditional prior using the last updated posterior as the proxy. $p(r_j^\tau \mid \gamma_t^\tau)$ is the likelihood of observing the feedback under prompt τ at step t.

With a temporal discounting factor $\lambda \in (0,1)$ applied for tracking the evolving policy [Qu et al., 2026a; Zhou et al., 2026], the posterior parameters are updated as:

$$\alpha_{t+1}^{\tau} = \lambda \cdot \alpha_t^{\tau} + (1-\lambda) \cdot \alpha_0^{\tau} + S_t^{\tau}, \qquad \beta_{t+1}^{\tau} = \lambda \cdot \beta_t^{\tau} + (1-\lambda) \cdot \beta_0^{\tau} + G - S_t^{\tau}. \tag{6}$$

Existing selection methods use this posterior to prioritize prompts with intermediate difficulty, as these are most likely to yield non-degenerate gradients [Bae et al., 2026; Qu et al., 2026a,b].

3.2 Composition Noise in GRPO

Previous online prompt selection methods model prompt difficulty using success rates and update it via posterior probabilities [Mao et al., 2026; Qu et al., 2026a,b; Zheng et al., 2025b; Zhou et al., 2026], while suffer from *composition noise* as illustrated in Fig. 1 (b). For an in-depth analysis, we formulate the explicit form of the aforementioned *composition noise*. The variance of the advantage for each response under the same prompt τ admits a precise decomposition via the law of total variance, expressed as:

$$\underbrace{\mathrm{Var}\big(\hat{A}_j^{\tau} \mid \gamma^{\tau}\big)}_{\text{total}} = \underbrace{\mathrm{Var}\big(\mathbb{E}[\hat{A}_j^{\tau} \mid r_j^{\tau}, \gamma^{\tau}] \,\big|\, \gamma^{\tau}\big)}_{\text{composition-free signal}} + \underbrace{\mathbb{E}\big[\mathrm{Var}(\hat{A}_j^{\tau} \mid r_j^{\tau}, \gamma^{\tau}) \,\big|\, \gamma^{\tau}\big]}_{\text{composition noise } \xi(\gamma^{\tau})}. \tag{7}$$

The first term describes the variation in advantage among different responses to the same prompt, which distinguishes correct responses from incorrect ones, reflecting the genuine signal. The second term describes that, even when the response is fixed as correct or incorrect, its advantage still varies across different groups — a variation that arises entirely from the random composition of the group. This forms the *composition noise* term $\xi(\gamma^{\tau})$, which is intrinsic and independent to prompt selection. However, $\xi(\gamma^{\tau})$ is strictly positive for all $\gamma^{\tau} \in (0,1)$ (Theorem D.1) and cannot be reduced by prompt selection alone. In standard GRPO, the realized advantage $\hat{A}_j^{\tau}$ directly enters the policy gradient, so the composition-noise term in Eq.(7) contributes variance to gradient estimation. Since prompt-selection methods assess prompt informativeness through sampled group rewards generated by the same policy, this noise also makes such assessments less faithful. Therefore, rather than using the noisy realized advantage $\hat{A}_j^{\tau}$, we target its conditional expectation $\mathbb{E}[\hat{A}_j^{\tau} \mid r_j^{\tau}, \gamma^{\tau}]$, which depends only on the response reward and the prompt's latent difficulty and removes the composition-noise term by construction.

4 Method

The decomposition in Eq. (7) identifies $\mathbb{E}[\hat{A}_j^{\tau} \mid r_j^{\tau}, \gamma^{\tau}]$ as the composition-free target. This section formalizes that target (§4.1), recovers it via posterior-predictive marginalization (§4.2), and extends the same principle to prompt selection (§4.3). In this section, we discuss our method with a given prompt τ. Hence, we omit superscript τ for simplicity.

4.1 The Intrinsic Advantage

The intrinsic advantage answers the question the standard GRPO advantage fails to: *given that this response is correct (or not) on a prompt of difficulty γ, what advantage should it receive on average?* Intuitively, it should depend only on the response reward and the prompt intrinsic difficulty, never on which peers happened to be drawn. Therefore, we first give the definition of the intrinsic advantage $a_j^{\star}(\gamma)$ and its closed form below.

Definition 1 (Intrinsic advantage). For a prompt τ with success rate $\gamma \in (0,1)$, the *intrinsic advantage* of response j is

$$a_j^{\star}(\gamma) \;:=\; \mathbb{E}\big[\hat{A}_j \,\big|\, r_j, \gamma\big], \tag{8}$$

Proposition 4.1 (Closed form). *Under* $r_j \mid \gamma \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(\gamma)$,

$$a_j^{\star}(\gamma) = \begin{cases} +\mu_G(\gamma), & r_j = 1, \\ -\mu_G(1-\gamma), & r_j = 0, \end{cases} \tag{9}$$

where

$$\mu_G(p) \;=\; \sum_{S=1}^{G-1} \binom{G-1}{S-1} p^{S-1}(1-p)^{G-S} \sqrt{\frac{G-S}{S}}, \; G \geq 2. \tag{10}$$

Note that in the degenerate cases where $S = \{0, G\}$ contribute zero and are omitted from the sum. Applying such intrinsic advantage $a_j^\star(\gamma)$ for $\hat{A}_j$ guarantees that $\mathrm{Var}[a_j^\star \mid r_j, \gamma] = 0$. Therefore, the total variance in Eq.(7) degenerates into the composition-free signal term alone and intrinsically eliminates the composition-noise term. However, $a_j^\star$ depends on the latent success rate γ, which can not be observed or explicitly derived. Therefore, we recover it through closed-form marginalization over the Beta posterior.

4.2 Marginalized Posterior-Predictive Advantage Denoising

To avoid each response's self-influence [Liu et al., 2025], we marginalize the latent success rate γ over a leave-one-out Beta posterior that excludes response j's own reward, formulated as:

$$p(\gamma \mid \mathcal{H}_t, r_{t;-j}) \;=\; \mathrm{Beta}(\alpha'_{t+1;j},\, \beta'_{t+1;j}), \tag{11}$$

$$\alpha'_{t+1;j} := \alpha_{t+1} - r_{t;j}, \;\; \beta'_{t+1;j} := \beta_{t+1} - 1 + r_{t;j}, \tag{12}$$

where group-shared α_{t+1} and β_{t+1} are calculated according to Eq. (6). $r_{t;-j}$ denotes the other rewards in the group except for j-th response at t-th iteration. This construction guarantees that the posterior used to estimate the advantage of response j is conditionally independent of r_j; consequently, the resulting estimator is free from bias induced by the response it evaluates. At each training step t, we marginalize $a_j^\star(\gamma)$ over the leave-one-out posterior to obtain a composition-free advantage estimate:

$$\hat{a}_j^{\mathrm{MaPP}} \;:=\; \mathbb{E}_{p(\gamma|\mathcal{H}_t, r_{t;-j})}\big[a_j^\star(\gamma)\big] \;=\; \int_0^1 a_j^\star(\gamma)\; p(\gamma \mid \mathcal{H}_t, r_{t;-j})\; \mathrm{d}\gamma. \tag{13}$$

With Eq.(13), we average all plausible values of γ, which is weighted by the observed optimization history and forms $\hat{a}_j^{\mathrm{MaPP}}$, called MaPP-AD. Consequently, the closed form of our tractable integral MaPP-AD is given by Beta–Binomial conjugacy.

Proposition 4.2 (Closed form)**.** *At training step t, let $\alpha'_j := \alpha'_{t+1;j}$ and $\beta'_j := \beta'_{t+1;j}$ for brevity. Then*

$$\hat{a}_j^{\mathrm{MaPP}} = \begin{cases} +\displaystyle\sum_{S=1}^{G-1} w_S(\alpha'_j,\, \beta'_j)\, \sqrt{\frac{G-S}{S}}, & r_j = 1, \\ -\displaystyle\sum_{S=1}^{G-1} w_S(\beta'_j,\, \alpha'_j)\, \sqrt{\frac{G-S}{S}}, & r_j = 0, \end{cases} \tag{14}$$

where $w_S(\alpha, \beta) := P_{\mathrm{BB}}(S-1 \mid G-1,\, \alpha, \beta) = \binom{G-1}{S-1} \dfrac{B(\alpha + S - 1,\, \beta + G - S)}{B(\alpha, \beta)}$ is the Beta-Binomial PMF evaluated at $S-1$ successes out of $G-1$ trials.

The weight $w_S(\alpha'_j, \beta'_j)$ has a concrete meaning: it is the posterior-predictive probability that a hypothetical fresh group of $G-1$ responses would contain $s-1$ correct ones. The estimator evaluates the advantage at every plausible composition s and averages under the posterior belief, replacing the single realized composition with its full predictive distribution. As the posterior concentrates around the true pass rate with more observations, $\hat{a}_j^{\mathrm{MaPP}}$ converges to the intrinsic advantage $a_j^\star(\gamma)$. We show in Appendix D that $\hat{a}_j^{\mathrm{MaPP}}$ achieves provably lower MSE than GRPO once the posterior accumulates a few updates per prompt.

4.3 Marginalized Posterior-Predictive Prompt Selection

§4.1–4.2 focused on denoising response-level advantage estimation. We now show how the same posterior-predictive principle also yields a more faithful prompt-selection score. Existing selection methods score prompts via point estimates of difficulty and select the top-k candidates closest to a target difficulty [Mao et al., 2026; Qu et al., 2026a; Zheng et al., 2025b; Zhou et al., 2026]. This point-estimate approach mirrors the same limitation exposed in §3.2: a single value of γ is substituted for the latent quantity, discarding posterior uncertainty.

We propose instead a principled informativeness score derived from the same marginalization principle. Define the probability that a prompt yields a non-degenerate gradient as $V_G(\gamma) = 1 -$

Table 1: Evaluation results on CountDown3to4 and mathematical reasoning benchmarks. **Bold** indicates the best result. CountDown is evaluated after CountDown3to4 training, while all other benchmarks use math-trained models. Rollouts and GPU hours report math-training cost.

Models	Methods	CountDown	AMC	MATH500	Minerva.	Olympiad.	Avg.↑	Rollouts↓	GPU Hours↓
Qwen3-4B	Random	80.95	54.22	78.63	28.31	38.25	56.07	563k	90h
	MoPPS	83.68	60.24	81.04	28.31	44.28	59.51	563k	81h
	DPS	83.42	59.04	80.65	28.68	43.98	59.15	563k	81h
	DS	83.83	60.24	82.06	30.88	45.03	60.41	2252k	209h
	Ours	**85.67**	**62.65**	**82.66**	**31.25**	**47.59**	**61.96**	**563k**	**81h**
Qwen3-8B	Random	81.29	55.42	81.05	30.51	42.17	58.09	563k	104h
	MoPPS	84.69	62.65	82.46	30.15	50.60	62.11	563k	88h
	DPS	85.80	60.24	81.45	29.78	48.49	61.15	563k	89h
	DS	83.79	61.45	82.86	**31.62**	48.34	61.61	2252k	230h
	Ours	**87.51**	**63.86**	**83.47**	30.88	**52.26**	**63.60**	**563k**	**88h**
R1-Distill-7B	Random	78.92	57.83	80.24	27.57	38.86	56.68	563k	63h
	MoPPS	80.75	59.20	81.45	26.84	41.11	57.87	563k	59h
	DPS	80.03	59.04	79.64	25.74	40.66	57.02	563k	58h
	DS	81.52	**62.65**	80.85	27.81	41.57	58.88	2231k	152h
	Ours	**82.24**	61.45	**82.26**	**28.31**	**42.17**	**59.29**	**563k**	**58h**

$(\gamma)^G - (1-\gamma)^G$, which peaks at $\gamma = 1/2$ and vanishes at the extremes. Since γ is latent, we marginalize against the posterior:

$$V_G^{\mathrm{MaPP}}(\alpha^\tau, \beta^\tau) \;:=\; \mathbb{E}_{\gamma\sim p(\gamma|\mathcal{H}^\tau)}\big[V_G(\gamma)\big] \;=\; 1 - \frac{B(\alpha^\tau{+}G,\, \beta^\tau)}{B(\alpha^\tau, \beta^\tau)} - \frac{B(\alpha^\tau,\, \beta^\tau{+}G)}{B(\alpha^\tau, \beta^\tau)}. \tag{15}$$

Since V_G is concave, Jensen's inequality gives $V_G^{\mathrm{MaPP}} \le V_G(\mathbb{E}[\gamma^\tau \mid \mathcal{H}^\tau])$: the marginalized score automatically down-weights prompts whose posterior is wide, avoiding wasted rollouts on prompts that appear informative under a point estimate but whose true difficulty remains uncertain. Batches are drawn by sampling prompts with probability proportional to $V_G^{\mathrm{MaPP}}(\alpha^\tau, \beta^\tau)$, preserving exploration over prompts whose posterior is still consolidating.

Unified posterior-predictive gradient. Combining posterior-predictive selection with the posterior-predictive advantage of §4.2, MaPP yields the following curriculum-weighted surrogate update:

$$g_{\mathrm{MaPP}} = \mathbb{E}_{\tau\sim q,\, \{y_j^\tau\}_{j=1}^G\sim\pi_\theta(\cdot|\tau)} \left[\frac{1}{G}\sum_{j=1}^{G} \hat{a}_j^{\mathrm{MaPP},\tau} \nabla_\theta \log \pi_\theta(y_j^\tau \mid \tau)\right], \quad q(\tau) \propto V_G^{\mathrm{MaPP}}(\alpha^\tau, \beta^\tau). \tag{16}$$

Here, both the batch distribution $q(\tau)$ and the per-response weight $\hat{a}_j^{\mathrm{MaPP},\tau}$ are posterior-predictive marginalizations against the same shared Beta posterior and are held fixed during each policy update. The history-dependent distribution q defines the training curriculum, without importance correction to the uniform-prompt objective in Eq. (1). Both components are modular: each can be adopted independently or composed with existing selection strategies, adding $O(|B| \cdot G^2)$ per-step overhead, negligible compared to rollout and backpropagation. Algorithm 1 gives the full training loop.

5 Experiments

5.1 Experimental Setup

We evaluate MaPP across three reasoning domains, mathematics, numerical planning, and visual geometry, using diverse model backbones to assess generality.

Tasks. For **Mathematics**, we train on MATH [Hendrycks et al., 2021] (7,500 problems) and evaluate on AMC23, MATH500 [Lightman et al., 2023], Minerva Math [Lewkowycz et al., 2022], and OlympiadBench [He et al., 2024]. For **Numerical planning**, we use the Countdown Number Game: train on Countdown-34 [Pan et al., 2025], evaluate on held-out CD-34 and harder CD-4 (four source numbers, larger search space). For **Visual geometry**, we train and evaluate on Geometry3k [Hiyouga, 2025; Lu et al., 2021], pairing diagrams with multi-step reasoning questions.

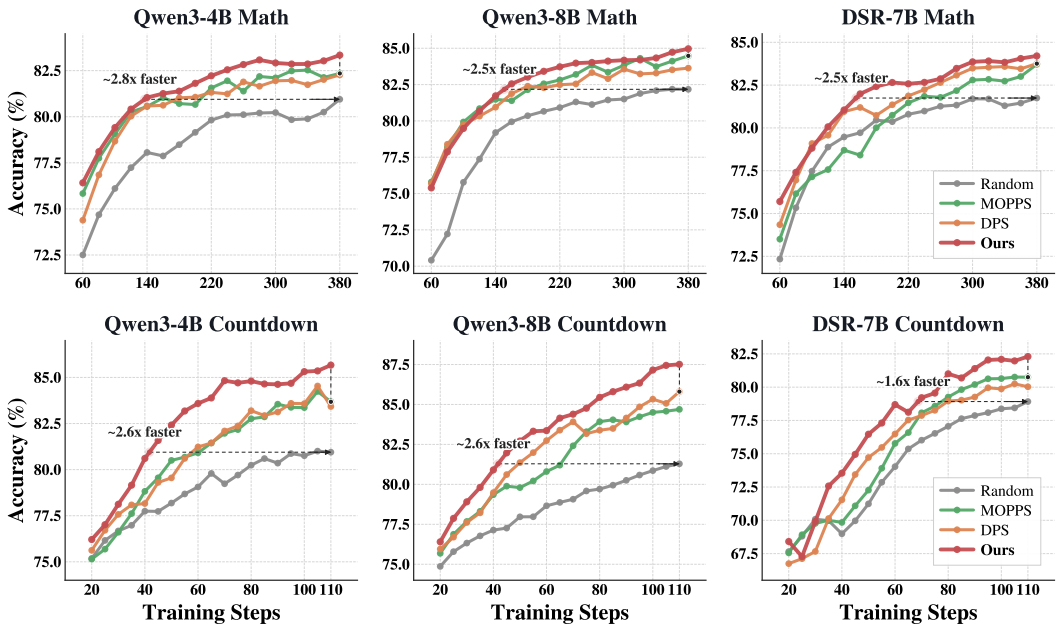


Figure 3: Performance comparisons of different methods and models on the MATH and Countdown tasks. Our proposed MaPP outperforms the existing SOTA methods and other baselines in both training efficiency and performance.

Models. For mathematics and planning, we use three text-only models: Qwen3-4B-Base, Qwen3-8B-Base [Yang et al., 2025], and DeepSeek-R1-Distill-Qwen-7B [Guo et al., 2025]. For visual geometry, we adopt Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct [Bai et al., 2025].

Baselines. We compare against four methods spanning online prompt selection paradigms [Qu et al., 2026b]. **(1) Random**: uniform sampling (standard GRPO baseline) [Shao et al., 2024]. **(2) MoPPS** [Qu et al., 2026a]: prediction-based; models success rate as Beta variable, uses Thompson sampling for intermediate difficulty. **(3) DPS** [Mao et al., 2026]: prediction-based; frames solving progress as a 3-state HMM, prioritizes partially-solved prompts. **(4) Dynamic Sampling (DS)** [Yu et al., 2025]: evaluation-based; oversamples $4\times$ batch, rolls out all, filters zero-variance prompts. Serves as compute-intensive oracle baseline.

5.2 Main Results

Mathematics. Tab. 1 reports evaluation results across five mathematics benchmarks for three model backbones. Across all scales, MaPP consistently achieves the highest average accuracy, outperforming both prediction-based methods (MoPPS, DPS) and the compute-intensive oracle baseline DS. On Qwen3-4B, MaPP obtains an average score of 61.96, improving over the strongest prediction-based baseline MoPPS by +2.45 and even surpassing DS by +1.55, while using only 25% of its rollout budget (563k vs. 2252k). Similar trends hold on Qwen3-8B (+1.49 over MoPPS) and R1-Distill-7B (+1.42 over MoPPS), confirming that the improvement is consistent across model families and scales.

Planning and visual geometry. On the Countdown task (Fig. 3, bottom row), MaPP achieves final accuracy gains of +1.84 (Qwen3-4B) and +1.71 (Qwen3-8B) over the strongest baseline at convergence. For visual geometry on Geometry3k (Fig. C(a)), MaPP outperforms the strongest baseline by +1.04 on Qwen2.5-VL-3B-Instruct and +2.40 on Qwen2.5-VL-7B-Instruct. These results demonstrate that MaPP generalizes beyond text-only mathematical reasoning to both planning and multi-modal tasks.

Training efficiency. Fig. 3 plots accuracy against training steps across all six model-task combinations. MaPP converges substantially faster than all baselines: on MATH, it reaches the same accuracy as Random Sample by approximately 2.5–2.8$\times$ faster convergence speed, while the acceleration rates range from 1.6$\times$ to 2.6$\times$ on Countdown dataset. MaPP introduces no additional rollout cost over

Table 2: Sensitivity to the temporal discount λ. Trained on Qwen3-4B, Countdown.

λ	CountDown3to4	CountDown4	Avg.
0.0	83.51	66.19	74.85
0.1	85.17	70.04	77.61
0.5	**86.76**	**72.24**	**79.50**
0.7	85.95	69.18	77.57
0.9	83.08	67.33	75.21
1.0	85.82	70.72	78.27

Table 3: Robustness to rollout group size G. Trained on Qwen3-4B, Countdown.

	$G = 4$			$G = 16$		
Method	CD-34	CD-4	#Rollouts	CD-34	CD-4	#Rollouts
Random	79.39	60.14	123k	82.54	64.01	492k
MoPPS	80.22	62.70	123k	86.39	72.79	492k
DPS	82.17	63.26	123k	86.90	70.18	492k
DS	81.85	64.11	490k	83.87	66.66	1855k
Ours	**82.45**	**64.44**	123k	**88.33**	**73.42**	492k

prediction-based methods, and requires less than half the runtime of DS across all settings (Tab. 1, last two columns).

Discussion. Several patterns emerge from the results. First, the gains are most pronounced on challenging benchmarks such as OlympiadBench (+3.31 on Qwen3-4B), suggesting that cleaner gradient signals from advantage denoising disproportionately benefit the learning of harder reasoning skills. Second, MaPP significantly improves over MoPPS and DPS despite sharing the same Beta posterior infrastructure, confirming that the gains stem from how the posterior is used, specifically marginalization for both advantage estimation and prompt selection, rather than from a better posterior itself. Third, the improvement persists across three task families and five model backbones, indicating that posterior-predictive marginalization provides a general refinement to GRPO learning signals rather than a task-specific heuristic.

5.3 Ablation Analysis

Effect of the two components. MaPP comprises two modules: the posterior-predictive advantage (MaPP-AD) and the posterior-predictive selection score (MaPP-PS). Tab. 4 disentangles the contributions of MaPP-AD and MaPP-PS on Qwen3-4B. Starting from standard GRPO, adding MaPP-PS alone yields +1.62 improvement on MATH, while MaPP-AD alone yields a larger +2.66. Combining both produces +3.44, exceeding the sum of individual gains, suggesting the two modules are synergistic. To verify complementarity with existing methods, we plug MaPP-AD into MoPPS and DPS without modifying their selection logic; it consistently improves both (+1.04 and +1.12 respectively), demonstrating that advantage denoising provides additional gains beyond prompt selection.

Table 4: **Ablation on the two components.** Trained on MATH, Qwen3-4B. MaPP-AD: posterior-predictive advantage (§ 4.2). MaPP-PS: posterior-predictive selection (§ 4.3).

Method	MATH	Olympiad.	Avg.
GRPO	79.41	38.25	58.83
+ MaPP-PS	81.03	43.17	62.10
+ MaPP-AD	82.07	44.52	63.30
+ AD + PS	**82.85**	**47.59**	**65.22**
MoPPS	82.17	44.28	63.23
+ MaPP-AD	**82.63**	**45.90**	**64.27**
DPS	82.24	43.98	63.11
+ MaPP-AD	**82.71**	**45.75**	**64.23**

Sensitivity to the temporal discount λ. The temporal discount λ controls the trade-off between tracking speed and posterior stability. Tab. 2 and Fig. C(b) reports MaPP performance on Countdown under varying λ. Performance peaks at λ=0.5 (79.50 avg.) and degrades at both extremes: λ=0.0 (74.85) forgets all history and reduces the posterior to single-step evidence, while λ=0.9 (75.21) retains stale observations that lag behind the evolving policy. The moderate sensitivity suggests that a reasonable balance between recency and stability is sufficient, and the optimal value does not require fine-grained tuning. We use λ=0.5 for all other experiments.

Robustness to group size. Tab. 3 and Fig. B evaluate MaPP on the Countdown benchmark under group sizes G=4 and G=16, complementing the default G=8 used on in main experiment. MaPP achieves the highest accuracy across both settings: at G=4, it outperforms the best baseline by +0.28 (CD-34) and +0.33 (CD-4); at G=16, the margin widens to +1.43 (CD-34) and +0.63 (CD-4). The wider margin at larger G is expected: more responses per group provide richer per-step evidence for the posterior, enabling more precise advantage corrections that outweigh the averaging effect of larger groups. This confirms that posterior-predictive marginalization scales favorably with group size.

Isolating scoring and sampling. To disentangle posterior marginalization from the sampling rule in MaPP-PS, we conduct a controlled 2×2 ablation, fixing MaPP-AD, the training data, rollout budget, and all other hyperparameters. We compare the posterior-mean plug-in score $V_G(\hat{\gamma})$, where $\hat{\gamma} = \alpha/(\alpha + \beta)$, with the marginalized score V_G^{MaPP}, each paired with top-B selection or proportional sampling. As shown in Table 5, marginalization improves accuracy by 1.97 and 1.11 percentage points under top-B and proportional sampling, respectively. Proportional sampling further improves the plug-in and marginalized scores by 1.12 and 0.26 points. Thus, the benefit of marginalization persists under both sampling rules, while combining marginalization with proportional sampling achieves the best result.

Table 5: Mechanism isolation in MaPP-PS. Accuracy (%) on the MATH test set using Qwen3-0.6B, trained for 75 steps with a batch size of 1024.

Selection score	Top-B	Proportional
Plug-in $V_G(\hat{\gamma})$	61.18	62.30
Marginalized V_G^{MaPP}	63.15	**63.41**

6 Conclusion

We identified composition noise in the GRPO advantage due to uncertainty in group composition, which degrades gradient estimation and prompt selection. To address this, we proposed MaPP: a unified framework that extends the per-prompt Beta posterior to a closed-form posterior-predictive advantage estimator with provably diminishing MSE and an uncertainty-aware selection score, addressing both data-efficiency losses through a single shared posterior. Experiments on mathematics, planning, and visual geometry with five backbones show consistent gains in accuracy and convergence over GRPO and strong baselines, with no additional rollout overhead.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grant No. 623B2016, the Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, and the Basic Research Cultivation Program of Zhongguancun Academy under Project No. C20250602. Discussions of the experiments were partially supported by the Ministry of Economic Development of the Russian Federation (agreement identifier 000000C313925P3X0002, Grant No. 139-15-2025-004 dated April 17, 2025).

References

S. Bae, J. Hong, M. Y. Lee, H. Kim, J. Nam, and D. Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 700–719, 2026.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Xu, M. Yan, J. Lin, D. Liu, B. Yu, B. Hui, B. Li, J. Zhou, J. Lin, C. Zhou, J. Zhou, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

G. Casella and C. P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

X. Chen, J. Lu, M. Kim, D. Zhang, J. Tang, A. Piché, N. Gontier, Y. Bengio, and E. Kamalloo. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*, 2025.

D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

C. He, R. Luo, Y. Bai, S. Hu, Z. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, 2024.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Hiyouga. Geometry3k: A large-scale multi-modal geometry reasoning dataset. `https://huggingface.co/datasets/hiyouga/geometry3k`, 2025.

A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Z. Lin, M. Lin, Y. Xie, and R. Ji. Cppo: Accelerating the training of group relative policy optimization-based reasoning models. *arXiv preprint arXiv:2503.22342*, 2025.

H. Liu, Y. Feng, Y. Mao, D. Zhou, J. Peng, and Q. Liu. Action-depedent control variates for policy optimization via stein's identity. *arXiv preprint arXiv:1710.11198*, 2017.

Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

P. Lu, R. Gong, S. Jiang, L. Qiu, S. Huang, X. Liang, and S.-C. Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021.

Y. Mao, Y. Qu, Q. Wang, H. Zou, and X. Ji. Dynamics-predictive sampling for active rl finetuning of large reasoning models. *arXiv preprint arXiv:2603.10887*, 2026.

J. Pan, J. Zhang, X. Wang, L. Yuan, H. Peng, and A. Suhr. Tinyzero. `https://github.com/Jiayi-Pan/TinyZero`, 2025. Accessed: 2025-01-24.

Y. Qu, Q. Wang, Y. Mao, V. T. Hu, B. Ommer, and X. Ji. Can prompt difficulty be online predicted for accelerating rl finetuning of reasoning models? In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1240–1250, 2026a.

Y. Qu, Q. Wang, Y. Mao, H. Zou, Y. Jiang, W. Liu, C. Bai, K. Yang, Y. Chen, S. Yang, et al. Small generalizable prompt predictive models can steer efficient rl post-training of large reasoning models. *arXiv preprint arXiv:2602.01970*, 2026b.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

Z. Stojanovski, O. Stanley, J. Sharratt, R. Jones, A. Adefioye, J. Kaddour, and A. Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards, 2025. URL `https://arxiv.org/abs/2505.24760`.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

F. Yang, Z. Chen, X. Wang, X. Lu, J. Chai, G. Yin, W. Lin, S. Ma, F. Zhuang, D. Wang, et al. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*, 2026.

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Y. Yue, Y. Yuan, Q. Yu, X. Zuo, R. Zhu, W. Xu, J. Chen, C. Wang, T. Fan, Z. Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Y. Zeng, Z. Sun, B. Ji, E. Min, H. Cai, S. Wang, D. Yin, H. Zhang, X. Chen, and J. Wang. Cures: From gradient analysis to efficient curriculum learning for reasoning llms. *arXiv preprint arXiv:2510.01037*, 2025.

R. Zhang, D. Arora, S. Mei, and A. Zanette. Speed-rl: Faster training of reasoning models via online curriculum learning. *arXiv preprint arXiv:2506.09016*, 2025.

C. Zheng, S. Liu, M. Li, X.-H. Chen, B. Yu, C. Gao, K. Dang, Y. Liu, R. Men, A. Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

H. Zheng, Y. Zhou, B. R. Bartoldson, B. Kailkhura, F. Lai, J. Zhao, and B. Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. *arXiv preprint arXiv:2506.02177*, 2025b.

H. Zheng, Y. Zhou, B. R. Bartoldson, B. Kailkhura, F. Lai, J. Zhao, and B. Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. *arXiv preprint arXiv:2506.02177*, 2025c.

X. Zhou, B. Zhu, H. Zhang, H. Wang, and Z. Guo. Efficient rlvr training via weighted mutual information data selection. *arXiv preprint arXiv:2603.01907*, 2026.

Appendix

A Algorithm

Algorithm 1 summarizes the complete MaPP training loop. At each step, prompts are first selected by sampling proportionally to the posterior-predictive score V_G^{MaPP} (§4.3). After rollout, the posterior-predictive advantage $\hat{a}_j^{\mathrm{MaPP}}$ replaces the standard GRPO advantage for policy update (§4.2). The shared Beta posterior is then updated from the new observations. The entire procedure adds only closed-form computations on top of standard GRPO, with no additional rollouts or model forward passes.

Algorithm 1 MAPP: Marginalized Posterior-Predictive Framework

Require: Prompt pool $\mathcal{T}$; policy π_θ; group size G; batch size $|\mathcal{B}|$; discount $\lambda \in (0,1)$; training steps T

Ensure: Updated policy π_θ

Initialize $(\alpha_0^\tau, \beta_0^\tau) \leftarrow (1,1)$ for all $\tau \in \mathcal{T}$ ▷ uninformative Beta prior

for $t = 0, \ldots, T-1$ **do**

▷ *Posterior-predictive prompt selection (§4.3)*

  for $\tau \in \mathcal{T}$ **do**

    $V_G^{\mathrm{MaPP}}(\alpha_t^\tau, \beta_t^\tau) \leftarrow 1 - \frac{B(\alpha_t^\tau + G, \beta_t^\tau)}{B(\alpha_t^\tau, \beta_t^\tau)} - \frac{B(\alpha_t^\tau, \beta_t^\tau + G)}{B(\alpha_t^\tau, \beta_t^\tau)}$ ▷ Eq. (15)

  end for

  Sample batch $\mathcal{B}_t \sim q(\tau) \propto V_G^{\mathrm{MaPP}}(\alpha_t^\tau, \beta_t^\tau)$

▷ *Rollout and reward collection*

  for $\tau \in \mathcal{B}_t$ **do**

    Draw $\{y_j^\tau\}_{j=1}^G \sim \pi_\theta(\cdot \mid \tau)$; observe rewards $\{r_{t;j}^\tau\}_{j=1}^G$

    $S_t^\tau \leftarrow \sum_{j=1}^G r_{t;j}^\tau$

  end for

▷ *Posterior-predictive advantage (§4.2)*

  for $\tau \in \mathcal{B}_t$, $j = 1, \ldots, G$ **do**

    $\alpha'^\tau_{t+1;j} \leftarrow \lambda \cdot \alpha_t^\tau + (1-\lambda) \cdot \alpha_0^\tau + S_t^\tau - r_{t;j}^\tau$ ▷ LOO posterior, Eq. (12)

    $\beta'^\tau_{t+1;j} \leftarrow \lambda \cdot \beta_t^\tau + (1-\lambda) \cdot \beta_0^\tau + G - S_t^\tau - 1 + r_{t;j}^\tau$

    if $r_{t;j}^\tau = 1$ **then**

      $\hat{a}_{t;j}^{\mathrm{MaPP},\tau} \leftarrow + \sum_{S=1}^{G-1} w_S(\alpha'^\tau_{t+1;j}, \beta'^\tau_{t+1;j}) \sqrt{(G-S)/S}$ ▷ Eq. (14)

    else

      $\hat{a}_{t;j}^{\mathrm{MaPP},\tau} \leftarrow - \sum_{S=1}^{G-1} w_S(\beta'^\tau_{t+1;j}, \alpha'^\tau_{t+1;j}) \sqrt{(G-S)/S}$

    end if

  end for

▷ *Policy update (Eq.* (16)*)*

  $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}(\theta)$ with $\nabla_\theta \mathcal{J} = \frac{1}{|\mathcal{B}_t|} \sum_{\tau \in \mathcal{B}_t} \frac{1}{G} \sum_{j=1}^G \hat{a}_{t;j}^{\mathrm{MaPP},\tau} \nabla_\theta \log \pi_\theta(y_j^\tau \mid \tau)$

▷ *Posterior update (Eq.* (6)*)*

  for $\tau \in \mathcal{B}_t$ **do**

    $\alpha_{t+1}^\tau \leftarrow \lambda \cdot \alpha_t^\tau + (1-\lambda) \cdot \alpha_0^\tau + S_t^\tau$

    $\beta_{t+1}^\tau \leftarrow \lambda \cdot \beta_t^\tau + (1-\lambda) \cdot \beta_0^\tau + G - S_t^\tau$

  end for

end for

B Implementation Details

B.1 Tasks and Datasets

Mathematics. We train on the MATH dataset [Hendrycks et al., 2021], which contains 7,500 competition-level problems spanning algebra, geometry, number theory, and combinatorics. Following prior work [Mao et al., 2026; Qu et al., 2026a,b], we evaluate on four benchmarks: AMC23, MATH500 [Lightman et al., 2023], Minerva Math [Lewkowycz et al., 2022], and OlympiadBench [He et al., 2024], using the datasets hosted by MOPPS [Qu et al., 2026a]. We adopt a binary reward

function following the default configuration in verl [Sheng et al., 2025]: a reward of 1 for correct and 0 otherwise.

Numerical planning. We adopt the Countdown Number Game [Pan et al., 2025], which requires combining given numbers using basic arithmetic operations to reach a target value. Training is conducted on a 2,000-problem subset of the Countdown-34 dataset. Evaluation uses two benchmarks: a 512-problem held-out split from Countdown-34 (CD-34), and a 512-problem subset from Countdown-4 (CD-4), a harder variant that provides four source numbers per problem and substantially enlarges the search space.

Visual geometry. We train on the 2,101-problem training split of the Geometry3k dataset [Hiyouga, 2025; Lu et al., 2021], which pairs geometric diagrams with multi-step reasoning questions. Evaluation is conducted on the official 601-problem test split.

B.2 Training Details

We adopt GRPO [Shao et al., 2024] as the default RLVR algorithm, implemented within the verl framework [Sheng et al., 2025]. At each training step, we sample $G{=}5$ responses per prompt for math tasks and $G{=}8$ responses per prompt for Countdown and Geometry, using temperature 1.0 and top-$p{=}1.0$ for rollout generation. We disable the KL penalty by setting $\beta{=}0$, consistent with Yu et al. [2025]. Training batch sizes are set to 256 for MATH and Countdown, with mini-batch sizes of 128 and 64 respectively, and 512 for Geometry3k with a mini-batch size of 256. The maximum response length is 1024 tokens for all tasks. Optimization is performed using AdamW [Loshchilov and Hutter, 2017] with learning rate 1×10^{-6}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and weight decay 0.01. We apply the Clip-Higher strategy from DAPO [Yu et al., 2025], which decouples clipping ranges with $\epsilon_{\text{low}}{=}0.2$ and $\epsilon_{\text{high}}{=}0.28$. All experiments are conducted on 8 NVIDIA H100 GPUs.

For all prediction-based prompt selection methods, the candidate pool size is set to $\hat{M}{=}8\times$ the training batch size B. For MoPPS [Qu et al., 2026a] and DPS [Mao et al., 2026], we follow their original top-B selection protocol, where prompts are ranked by the corresponding scoring criterion and the top-B prompts are selected. In contrast, MaPP does not perform deterministic top-B selection. Instead, each candidate prompt is assigned a retention probability $q(\tau) \propto V_G^{\mathrm{MaPP}}(\alpha^\tau, \beta^\tau)$, and the training batch is sampled according to these probabilities. For MaPP, the Beta prior is initialized as $(\alpha_0, \beta_0) = (1, 1)$ and the temporal discount is set to $\lambda{=}0.5$ across all tasks. For MoPPS [Qu et al., 2026a], we follow the original configuration with Beta prior $(\alpha_0, \beta_0) = (1, 1)$, target success probability $\gamma^*{=}0.5$, and decay factor $\lambda{=}0.5$. For DPS [Mao et al., 2026], we adopt the default three-state HMM with Dirichlet prior $\alpha_0 = (1, 1, 1)$ and decay ratio $\lambda{=}0.5$. For Dynamic Sampling [Yu et al., 2025], we use the verl implementation, which over-samples a candidate batch $4\times$ the training batch size and post-hoc filters out prompts with zero reward standard deviation.

C Additional Experiments

C.1 Empirical Calibration of MaPP-AD and MaPP-PS

We examine the empirical calibration of the intrinsic advantage target $a_j^\star(\gamma)$ and the posterior-predictive selection score V_G^{MaPP}. For MaPP-AD, we validate its intrinsic advantage target using a mid-training checkpoint on Countdown (Qwen3-4B, $G{=}8$), sampling 100 prompts and drawing 50 independent groups per prompt. Figure A(a) shows that the theoretical intrinsic advantage closely tracks the empirical mean GRPO advantage, with nearly all points falling along the ideal calibration line. The ±1 std band visualizes the variability of GRPO advantages induced by random group composition. For MaPP-PS, we collect the predicted $V_G^{\mathrm{MaPP}}(\alpha^\tau, \beta^\tau)$ and the empirical fraction of groups with non-zero reward variance at each training step, bin the observations by the predicted score, and compare. Figure A(b) shows that the predicted score tracks the empirical fraction within a ±0.05 band across the evaluated range. Together, these plots support the calibration of the intrinsic advantage target and the prompt-selection score.

C.2 Training Dynamics

Robustness to group size. Figure B shows accuracy versus training steps under group sizes $G{=}4$ and $G{=}16$ on Countdown. MaPP reaches the same accuracy as the strongest baseline by

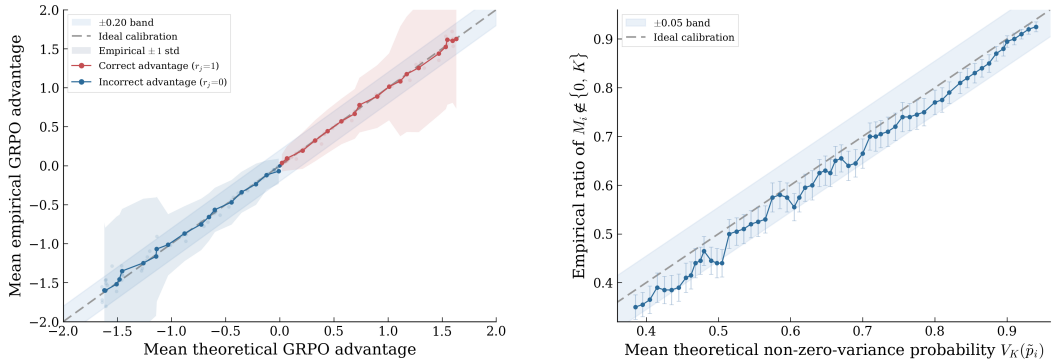


Figure A: **(a)** Calibration of the intrinsic advantage: the theoretical target $a_j^\star(\gamma)$ (Proposition 4.1) closely matches the empirical mean GRPO advantage across 100 prompts $\times$ 50 groups (G=8). The ± 1 std band reflects the variability of GRPO advantages induced by random group composition. **(b)** Calibration of the selection score: the posterior-predictive probability of a non-degenerate reward group, V_G^{MaPP} (Eq. (15)), tracks the empirical fraction of groups with non-zero reward variance within a ± 0.05 band.

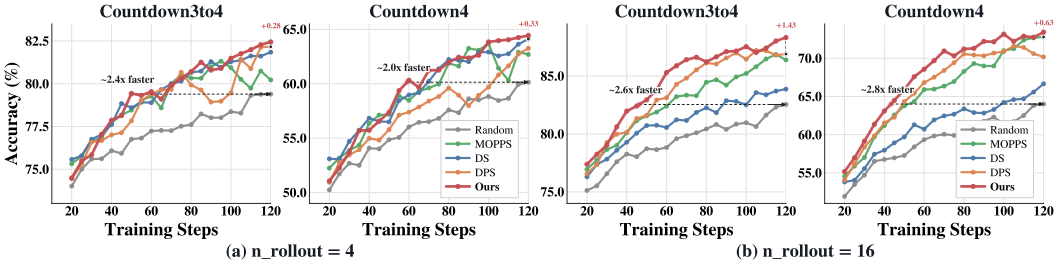


Figure B: Ablation experiments of our proposed MaPP method under different numbers of rollouts ($n = 4$ and $n = 16$, with $n = 8$ evaluated in the main experiments). The results show that MaPP performs consistently well across different rollout group sizes.

approximately 2.0–2.4$\times$ faster convergence speed at G=4, while the acceleration increases to 2.6–2.8$\times$ at G=16. The more pronounced speedup at larger G is consistent with the observation in §5.3 that richer per-step evidence enables sharper advantage corrections.

Visual geometry and sensitivity to λ. Figure C(a) shows training curves for the visual geometry task on Qwen2.5-VL-3B and Qwen2.5-VL-7B. MaPP outperforms the strongest baseline by +1.04 and +2.40 respectively at convergence, reaching comparable accuracy by approximately 1.4–1.5$\times$ faster convergence speed, confirming that MaPP generalizes to multi-modal settings. Figure C(b) shows training curves under varying λ on Countdown. λ=0.5 achieves the best final accuracy and convergence speed, while extreme values (λ=0.0 and λ=1.0) converge noticeably slower.

D Theoretical Analysis

All proofs in this section are for a single prompt τ at a fixed training step t; we drop the superscript τ for brevity and write $\alpha'_j := \alpha'_{t+1;j}$, $\beta'_j := \beta'_{t+1;j}$ following Proposition 4.2. The latent pass rate $\gamma \in (0, 1)$ is treated as a fixed (but unknown) parameter; expectations $\mathbb{E}[\cdot \mid \gamma]$ are taken over the rollout randomness only, while Bayesian quantities are computed under the leave-one-out posterior $\gamma \mid \mathcal{H}_t, r_{t;-j} \sim \mathrm{Beta}(\alpha'_j, \beta'_j)$. Under binary rewards, the GRPO advantage reduces to

$$\hat{A}_j = h(r_j, S) = \begin{cases} +\sqrt{(G-S)/S}, & r_j = 1, \\ -\sqrt{S/(G-S)}, & r_j = 0, \end{cases} \tag{17}$$

with $\hat{A}_j := 0$ when $S \in \{0, G\}$; correspondingly, both $a_j^\star(\gamma)$ and $\hat{A}_j$ drop the degenerate terms, so $\mathbb{E}[\hat{A}_j \mid r_j, \gamma] = a_j^\star(\gamma)$ for $\gamma \in (0, 1)$.

Theorem D.1 (Composition-noise lower bound)**.** *For any $G \geq 2$ and $\gamma \in (0,1)$,*

$$\mathbb{E}\big[(\hat{A}_j - a_j^\star(\gamma))^2 \,\big|\, \gamma\big] \;=\; \xi(\gamma) \;>\; 0, \tag{18}$$

where $\xi(\gamma)$ depends only on (γ, G) and is independent of the posterior or training history.

Proof. By Definition 1, $a_j^\star(\gamma) = \mathbb{E}[\hat{A}_j \mid r_j, \gamma]$. Conditioning on r_j and γ, the only remaining randomness comes from the other $G-1$ responses, so

$$\mathbb{E}\big[(\hat{A}_j - a_j^\star(\gamma))^2 \mid \gamma\big] = \mathbb{E}_{r_j}\big[\,\mathrm{Var}(\hat{A}_j \mid r_j, \gamma)\,\big|\,\gamma\big] =: \xi(\gamma). \tag{19}$$

Letting $k \sim \mathrm{Bin}(G-1, \gamma)$ denote the number of correct responses among the other $G-1$,

$$\xi(\gamma) = \gamma\;\mathrm{Var}_k\!\left[\sqrt{\tfrac{G-1-k}{k+1}}\right] + (1-\gamma)\;\mathrm{Var}_k\!\left[\sqrt{\tfrac{k}{G-k}}\right]. \tag{20}$$

The map $S \mapsto \sqrt{(G-S)/S}$ is strictly monotone on $\{1, \ldots, G-1\}$ and the binomial is non-degenerate for $\gamma \in (0,1)$, hence $\xi(\gamma) > 0$. Since ξ depends only on (γ, G), this lower bound is structural and cannot be reduced by any prompt-level posterior or selection scheme. □

Theorem D.2 (Frequentist MSE bound for MaPP)**.** *Fix $\gamma \in (0,1)$ and $G \geq 2$. Let $\tilde{\gamma} := \alpha'_j/(\alpha'_j + \beta'_j)$ and $n := \alpha'_j + \beta'_j$. There exist constants $C_1(\gamma, G), C_2(\gamma, G)$ depending only on (γ, G) such that*

$$\mathbb{E}\big[(\hat{a}_j^{\mathrm{MaPP}} - a_j^\star(\gamma))^2 \,\big|\, \gamma\big] \;\leq\; \underbrace{C_1\,(\tilde{\gamma} - \gamma)^2}_{\textit{calibration bias}} \;+\; \underbrace{\frac{C_2}{n+1}}_{=O(G/n)}. \tag{21}$$

Proof. Decompose the frequentist MSE under the rollout history $\mathcal{H}_t$ into squared bias and variance:

$$\mathbb{E}\big[(\hat{a}_j^{\mathrm{MaPP}} - a_j^\star(\gamma))^2 \mid \gamma\big] \;=\; \underbrace{\big(\mathbb{E}_{\mathcal{H}_t}[\hat{a}_j^{\mathrm{MaPP}} \mid \gamma] - a_j^\star(\gamma)\big)^2}_{\text{bias}^2} \;+\; \underbrace{\mathrm{Var}_{\mathcal{H}_t}(\hat{a}_j^{\mathrm{MaPP}} \mid \gamma)}_{\text{variance}}. \tag{22}$$

Since $a_j^\star = \pm\mu_G$ is a degree-$(G-1)$ polynomial with binomial coefficients, its first two derivatives are bounded on compact subsets of $(0,1)$:

$$L_G \;:=\; \sup_{\gamma \in K} \big|(a_j^\star)'(\gamma)\big| \;=\; O(G), \qquad M_G \;:=\; \sup_{\gamma \in K} \big|(a_j^\star)''(\gamma)\big| \;=\; O(G^2), \qquad K \Subset (0,1), \tag{23}$$

where the constants depend on the distance of K to $\{0,1\}$ and are bounded over the non-degenerate regime enforced by MaPP-PS via V_G^{MaPP}. A second-order Taylor expansion of $a_j^\star$ around γ gives, for γ' in a neighborhood of γ,

$$a_j^\star(\gamma') \;=\; a_j^\star(\gamma) + (a_j^\star)'(\gamma)\,(\gamma' - \gamma) + R(\gamma', \gamma), \qquad |R(\gamma', \gamma)| \leq \tfrac{1}{2} M_G\,(\gamma' - \gamma)^2. \tag{24}$$

Taking expectation under the leave-one-out posterior $\gamma' \sim \mathrm{Beta}(\alpha'_j, \beta'_j)$ at fixed $\mathcal{H}_t$, and using $\mathbb{E}_{\gamma'}[\gamma'] = \tilde{\gamma}$,

$$\hat{a}_j^{\mathrm{MaPP}} \;=\; a_j^\star(\gamma) + (a_j^\star)'(\gamma)\,(\tilde{\gamma} - \gamma) + \mathbb{E}_{\gamma'}[R(\gamma', \gamma)]. \tag{25}$$

Bias. Taking expectation of Eq. (25) over $\mathcal{H}_t$ at fixed γ and using $|\mathbb{E}_{\gamma'}[R]| \leq \frac{1}{2} M_G \mathbb{E}_{\gamma'}[(\gamma' - \gamma)^2]$,

$$\big|\mathbb{E}_{\mathcal{H}_t}[\hat{a}_j^{\mathrm{MaPP}} \mid \gamma] - a_j^\star(\gamma)\big| \;\leq\; L_G\,\big|\mathbb{E}_{\mathcal{H}_t}[\tilde{\gamma} \mid \gamma] - \gamma\big| + O\big(M_G/n\big), \tag{26}$$

where the last term absorbs both $\mathbb{E}_{\gamma'}[(\gamma' - \tilde{\gamma})^2] = \mathrm{Var}_{\gamma'}(\gamma') = O(1/n)$ and $(\tilde{\gamma} - \gamma)^2$. Squaring, the bias contributes $C_1\,(\tilde{\gamma} - \gamma)^2 + O(G^2/n^2)$ with $C_1 = L_G^2 = O(G^2)$; the higher-order term is dominated by the variance term below for $n \geq 1$.

Variance. The variance term equals the posterior variance of $a_j^\star(\gamma')$ under $\gamma' \sim \mathrm{Beta}(\alpha'_j, \beta'_j)$. Substituting Eq. (24) and using $\mathrm{Var}_{\gamma'}(\text{constants}) = 0$,

$$\mathrm{Var}_{\gamma'}\big(a_j^\star(\gamma')\big) \;\leq\; \big[(a_j^\star)'(\gamma)\big]^2\,\mathrm{Var}_{\gamma'}(\gamma') + O\big(M_G^2/n^2\big) \;\leq\; \frac{L_G^2}{4(n+1)} + O(1/n^2), \tag{27}$$

using $\mathrm{Var}_{\gamma'}(\gamma') = \alpha'_j\beta'_j/[n^2(n+1)] \leq 1/[4(n+1)]$. This gives $C_2 = L_G^2/4 = O(G)$.

Combining bias and variance yields Eq. (21).

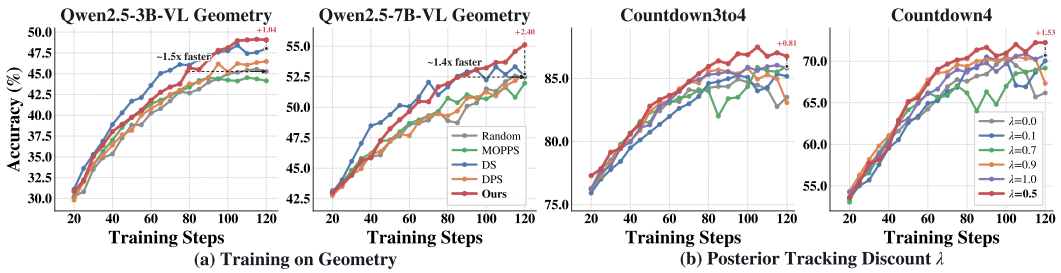


Figure C: (a): Performance of our method trained on the Qwen2.5-VL-Instruct series models and evaluated on the Geometry test set. (b): Self-ablation of MaPP under different posterior tracking discounts. The best performance is achieved at $\lambda = 0.5$ (Ours).

Corollary D.3 (Crossover)**.** *Combining Theorems D.1 and D.2, there exists a threshold*

$$n^\star(\gamma, G) \;\le\; \frac{C_2}{\xi(\gamma) - C_1(\tilde{\gamma} - \gamma)^2} - 1, \tag{28}$$

such that for all $n \ge n^\star$,

$$\mathbb{E}\big[(\hat{a}_j^{\mathrm{MaPP}} - a_j^\star(\gamma))^2 \,\big|\, \gamma\big] \;\le\; \xi(\gamma) \;=\; \mathbb{E}\big[(\hat{A}_j - a_j^\star(\gamma))^2 \,\big|\, \gamma\big], \tag{29}$$

i.e., MaPP strictly improves over GRPO at the per-response level. In practice, this threshold is small under standard initialization $(\alpha_0, \beta_0) = (1, 1)$ and is reached after a few posterior updates per prompt, ensuring that MaPP outperforms GRPO throughout the main training phase.

Connection to gradient estimation. The per-response MSE bounds in Theorems D.1 and D.2 transfer to the policy gradient under a standard score-function moment condition. Let $s_j := \nabla_\theta \log \pi_\theta(y_j \mid \tau)$ and let $g := \frac{1}{G}\sum_{j=1}^G \hat{a}_j s_j$, $g^\star := \frac{1}{G}\sum_{j=1}^G a_j^\star(\gamma) s_j$ denote the policy gradient and its oracle counterpart, where $\hat{a}_j$ denotes either $\hat{A}_j$ (GRPO) or $\hat{a}_j^{\mathrm{MaPP}}$ (MaPP), with per-response error $\epsilon_j := \hat{a}_j - a_j^\star(\gamma)$. Suppose $\mathbb{E}[\|s_j\|^2 \mid \tau] \le B^2$, as is routinely assumed in policy-gradient analysis. A direct application of Cauchy–Schwarz to $g - g^\star = \frac{1}{G}\sum_{j=1}^G \epsilon_j s_j$, together with symmetry across responses, yields

$$\mathbb{E}\big[\|g - g^\star\|^2 \,\big|\, \gamma\big] \;\le\; B^2 \cdot \mathbb{E}\big[\epsilon_1^2 \,\big|\, \gamma\big]. \tag{30}$$

Since this bound is monotone in the per-response MSE, the bounds of Theorems D.1 and D.2 immediately yield corresponding gradient-level bounds, and Corollary D.3 likewise transfers: MaPP achieves a smaller gradient MSE upper bound than GRPO whenever $n \ge n^\star$. Empirically, Fig. 1(b) and Fig. A show that composition noise produces visible variability in realized advantages across groups, which MaPP suppresses.

E Derivation of Closed Forms

All derivations in this section are for a single prompt τ; we drop the superscript τ for brevity and write $\alpha_j' := \alpha_{t+1;j}'$, $\beta_j' := \beta_{t+1;j}'$ following Proposition 4.2.

Intrinsic advantage (Proposition 4.1). Consider $r_j = 1$; the case $r_j = 0$ follows by symmetry. Conditional on $r_j = 1$ and γ, the remaining $G - 1$ responses each follow independent Bernoulli(γ) trials, so the total success count satisfies $S \in \{1, \dots, G\}$ with $\mathbb{P}(S = s \mid r_j = 1, \gamma) = \binom{G-1}{s-1}\gamma^{s-1}(1-\gamma)^{G-s}$. Since $\hat{A}_j = 0$ when $S = G$, taking expectation gives

$$a_j^\star(\gamma) = \mathbb{E}[\hat{A}_j \mid r_j{=}1,\, \gamma] = \sum_{S=1}^{G-1} \binom{G-1}{S-1} \gamma^{S-1}(1{-}\gamma)^{G-S} \sqrt{\frac{G{-}S}{S}} = \mu_G(\gamma). \tag{31}$$

For $r_j = 0$, by symmetry $a_j^\star(\gamma) = -\mu_G(1{-}\gamma)$.

Posterior-predictive advantage (Proposition 4.2). Marginalizing $a_j^\star(\gamma) = \mu_G(\gamma)$ against the leave-one-out posterior $\gamma \sim \mathrm{Beta}(\alpha'_j, \beta'_j)$ and exchanging summation with integration:

$$\begin{aligned}
\hat{a}_j^{\mathrm{MaPP}} &= \int_0^1 \mu_G(\gamma)\, \frac{\gamma^{\alpha'_j-1}(1-\gamma)^{\beta'_j-1}}{B(\alpha'_j,\beta'_j)}\,\mathrm{d}\gamma \\
&= \sum_{S=1}^{G-1} \binom{G-1}{S-1} \sqrt{\frac{G-S}{S}}\, \frac{1}{B(\alpha'_j,\beta'_j)} \underbrace{\int_0^1 \gamma^{\alpha'_j+S-2}(1-\gamma)^{\beta'_j+G-S-1}\,\mathrm{d}\gamma}_{=B(\alpha'_j+S-1,\ \beta'_j+G-S)} \\
&= \sum_{S=1}^{G-1} \underbrace{\binom{G-1}{S-1} \frac{B(\alpha'_j+S-1,\ \beta'_j+G-S)}{B(\alpha'_j,\beta'_j)}}_{=\,w_S(\alpha'_j,\beta'_j)\,=\,P_{\mathrm{BB}}(S-1|G-1,\alpha'_j,\beta'_j)} \sqrt{\frac{G-S}{S}}.
\end{aligned} \tag{32}$$

The case $r_j = 0$ follows by the change of variable $\gamma' = 1-\gamma$, which swaps α'_j and β'_j.

Selection score (Eq. (15)). The selection score operates across prompts, so we restore the superscript τ. By linearity, $V_G^{\mathrm{MaPP}} = 1 - \mathbb{E}_{\gamma\sim\mathrm{Beta}(\alpha^\tau,\beta^\tau)}[\gamma^G] - \mathbb{E}[(1-\gamma)^G]$. The first moment evaluates as

$$\mathbb{E}[\gamma^G] = \frac{1}{B(\alpha^\tau,\beta^\tau)} \int_0^1 \gamma^{\alpha^\tau+G-1}(1-\gamma)^{\beta^\tau-1}\,\mathrm{d}\gamma = \frac{B(\alpha^\tau+G,\ \beta^\tau)}{B(\alpha^\tau,\beta^\tau)}, \tag{33}$$

and $\mathbb{E}[(1-\gamma)^G]$ follows by symmetry, giving Eq. (15).

F Extension to Multi-Level Rewards

The main text presents MaPP under binary rewards for notational clarity, but the posterior-predictive marginalization principle extends naturally to discrete multi-level rewards $r_j^\tau \in \{v_1, \ldots, v_L\}$ by replacing the Beta–Binomial conjugate pair with the Dirichlet–Multinomial pair. The latent difficulty becomes a probability vector $\boldsymbol{\gamma}^\tau \in \Delta^{L-1}$ with a Dirichlet posterior, and the posterior-predictive advantage estimator is obtained by enumerating leave-one-out peer-count vectors weighted by their Dirichlet–Multinomial predictive probabilities. For moderate G and L, this involves $\binom{G+L-2}{L-1}$ terms (e.g., 36 for G=8, L=3), which is computationally negligible.

We validate this extension on Reasoning Gym's Shortest Path task [Stojanovski et al., 2025] using Qwen3-4B trained for 120 steps. A BFS-based verifier assigns three reward levels: invalid or no progress, strictly reduced distance to the goal, and an optimal solution or a correct infeasibility decision. All methods use the same reward scheme, and accuracy counts only outcomes receiving the highest reward. As shown in Table A, MaPP achieves the highest accuracy across all splits, improving average accuracy over the strongest baseline by 7.1 percentage points.

Table A: Accuracy (%) on Shortest Path with three-level rewards (Qwen3-4B, 120 training steps).

Method	OOD	ID	Dev	Avg.
GRPO	20.8	35.8	35.6	30.7
MoPPS	22.6	36.2	36.8	31.9
DPS	21.8	37.0	39.4	32.7
MaPP	**25.2**	**48.2**	**46.0**	**39.8**

G Data Examples

We provide illustrative data examples for each task below. Prompt templates for MATH and Geometry3k are adopted from the verl framework [Sheng et al., 2025], and the Countdown template follows Pan et al. [2025].

MATH Data Example

Prompt:

You have seven bags of gold coins. Each bag has the same number of gold coins. One day, you find a bag of 53 coins. You decide to redistribute the number of coins you have so that all eight bags you hold have the same number of coins. You successfully manage to redistribute all the coins, and you also note that you have more than 200 coins. What is the smallest number of coins you could have had before finding the bag of 53 coins?

Let's think step by step and output the final answer within `\boxed{}`.

Ground-Truth Answer:

203

Countdown Data Example

Prompt:

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer.

User: Using the numbers $[79, 8, 27, 47]$, create an equation that equals 91. You can use basic arithmetic operations $(+, -, \times, \div)$, and each number can only be used once. Show your work in `<think> </think>` tags, and return the final answer in `<answer> </answer>` tags, for example `<answer>`$(1 + 2)/3$`</answer>`.

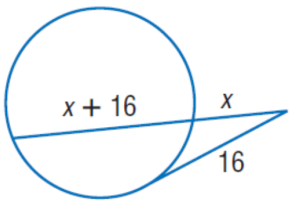
Assistant: Let me solve this step by step.
`<think>`

Ground-Truth Answer:

$79 + 47 - 27 - 8 = 91$

Geometry3K Data Example

Prompt:



Find x. Round to the nearest tenth if necessary. Assume that segments that appear to be tangent are tangent.

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{}`.

Ground-Truth Answer:

8

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state MaPP's main contributions: identifying composition noise, proposing posterior-predictive advantage denoising, and improving prompt selection with a shared Beta posterior. These claims are aligned with the theoretical analysis and empirical results reported in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: We discuss the scope of MaPP's Bernoulli–Beta formulation and its current focus on binary verifiable rewards. Appendix F further discusses how the posterior-predictive idea can be extended to multi-valued reward settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main theoretical results are stated with their assumptions in the paper, and the complete derivations and proofs are provided in the appendix. We also cross-reference the relevant propositions and theorems from the main text to the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the main algorithm, datasets, model backbones, baselines, rollout budgets, training hyperparameters, and compute resources needed to reproduce the experiments. Additional implementation details are included in the appendix to support reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code will be released publicly upon acceptance of the paper. All implementation details, hyperparameters, datasets, and compute resources are provided in Appendix B to support reproduction of the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the datasets, evaluation splits, model backbones, baselines, rollout budgets, optimizer, and key training hyperparameters in the main paper and Appendix B. These details are sufficient to understand and reproduce the reported experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Test accuracy is reported as average pass@1 over 16 independent generations per problem, which helps reduce evaluation variance. We also evaluate MaPP across multiple tasks, model backbones, and baselines to assess the consistency of the observed improvements.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute resources used for the experiments, including 8 NVIDIA H100 GPUs and the corresponding training time/GPU hours. These details are provided in the experimental setup and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics and uses publicly available datasets and standard evaluation protocols. We do not involve human-subject experiments, private data, or personally identifiable information.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper studies a data-efficient RLVR training method for reasoning models, rather than a new deployment scenario or application. Thus, it does not introduce direct societal impacts beyond the general considerations associated with the underlying language models.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release new pretrained models, image generators, or scraped datasets with high misuse risk. The experiments are conducted using existing models and public benchmarks, so no additional release safeguards are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the original papers and sources for the datasets, models, and code frameworks used in our experiments. We use publicly available assets and follow their corresponding licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce new datasets, model checkpoints, or benchmarks as new assets at submission time. The proposed method is fully described in the paper and appendix, and the code will be released upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects. Therefore, participant instructions, screenshots, and compensation details are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects. Therefore, IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the LLMs used as policy models in the RLVR experiments, including the model backbones, training setup, and evaluation protocols. Any LLM-based writing assistance, if used, did not affect the core methodology, experiments, or conclusions.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.