# BLOO: BAYESIAN LEAVE-ONE-OUT POLICY OPTIMIZATION FOR GROUP CREDIT ASSIGNMENT

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

## ABSTRACT

In group-based policy optimization, *e.g.*, GRPO, relative credit assignment relies on contrasting sibling responses. When rollouts yield identical binary outcomes in early or difficult tasks, standard empirical normalization collapses, leaving entire groups with zero gradient signal. Yet agreement among a few observed rewards does not mean the policy has no room to improve. We present BLOO, a simple RLVR framework that treats peer outcomes as finite evidence rather than absolute baselines. By computing a Jeffreys-smoothed sibling reference and scaling credit by the predictive standard deviation of the next outcome, BLOO maintains well-behaved, non-zero credit even in homogeneous groups. Beyond ensuring non-zero updates, we provide a comprehensive theoretical characterization of the retained signal: (i) we prove that the expected update follows a well-defined transformed objective; (ii) we establish that homogeneous groups account for at most $1/(2K)$ of the per-prompt mean signal coefficient; and (iii) we decompose gradient variance into cross-reward and within-reward components, revealing that credit retention does not guarantee variance reduction. Ablation analyses confirm the theoretical predictions, while comprehensive experiments on Qwen3-VL show that BLOO outperforms GRPO in both final and late-stage average validation accuracy.

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves reasoning policies using verifiable answers (DeepSeek-AI et al., 2025; Wang et al., 2025), and Group Relative Policy Optimization (GRPO) makes this approach practical by assigning relative credit to responses to the same prompt without a separate value network (Shao et al., 2024). The sampled group provides both the reference against which each response is judged and the scale of its update (Liu et al., 2025; Ge et al., 2026), so the usefulness of this feedback depends on how a small set of binary outcomes is interpreted.

The limitation becomes apparent when all sampled responses receive the same reward, as shown on the GRPO side of Figure 1(a), where five failures yield zero empirical mean and scale and therefore leave every response with zero credit (He et al., 2026). Excluding each response from the empirical mean still leaves only failures (Ahmadian et al., 2024), although five unsuccessful attempts

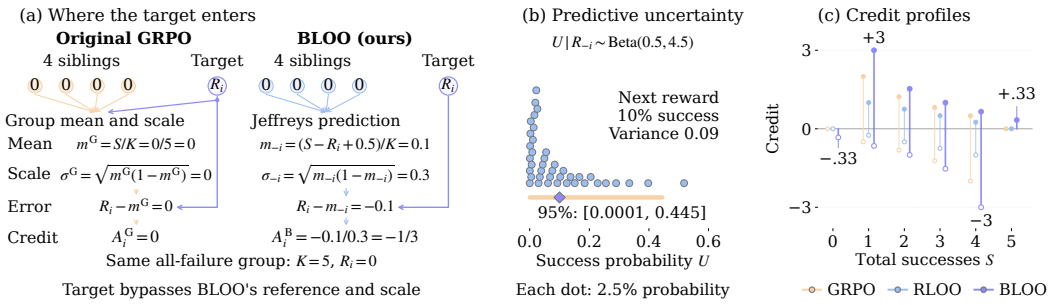


Figure 1: Empirical versus predictive credit at $K = 5$. Analytic comparisons show (a) target routing and all-failure credit, (b) posterior uncertainty after four failed siblings, and (c) credits across group success counts. In figure (b), each dot carries 2.5% posterior mass, orange marks the central 95%, and the diamond marks the predictive mean; filled/open markers in figure (c) denote success/failure.



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

cannot establish that a correct answer is impossible. Such *homogeneous groups*, *i.e.*, groups whose rewards all agree, occur $59.05\%$ of the time for five responses from a policy with a $10\%$ success rate, making the loss of credit a frequent consequence of limited sampling. In visual reasoning, identical failure labels can also conceal different perceptual and reasoning errors (Zhang et al., 2024), further motivating a comparison that uses the observed agreement without treating it as certainty about future outcomes. Existing approaches change the available evidence or its interpretation, with DAPO replacing homogeneous groups and GRESO filtering prompts before rollout (Yu et al., 2025; Zheng et al., 2025b). Statistical approaches augment normalization with virtual rewards or a Beta model (He et al., 2026; Xiao et al., 2025), shrink estimates using global statistics as in EBPO (Han et al., 2026), or accumulate prompt-specific evidence (Kim et al., 2026). AERO combines Bayesian estimation with adaptive sampling (Zhang et al., 2026), while richer quality feedback can distinguish equally rewarded trajectories in vision-language-action tasks (Deng et al., 2026). These approaches motivate studying how a fixed group's binary rewards can provide a useful reference for each response.

To construct this reference, we introduce BLOO, Bayesian Leave-One-Out Optimization, which predicts each target response's reward using the other responses, referred to as its *siblings*. As shown in Figure 1(a), GRPO includes the target reward in its own reference and scale, whereas BLOO derives both from siblings and uses the target reward only to compute the prediction error. A Jeffreys prior keeps the sibling success estimate away from zero and one, while the predictive standard deviation of the next binary reward determines the weight assigned to a deviation from that estimate.

As shown in Figure 1(b), four failed siblings remain compatible with a range of success probabilities under the working Bayesian model, yielding a predicted success probability of $0.1$ and a next-reward standard deviation of $0.3$. Figure 1(c) extends this comparison across all possible group success counts, showing that BLOO retains finite, nonzero credits at both homogeneous endpoints while adjusting their magnitudes to the available evidence. Importantly, computing these credits requires only $O(K)$ scalar work for $K$ responses, without additional rollouts, a fitted critic, or reward history.

For the value of retained credit assessment, we must determine how much learning signal it supplies and how much noise accompanies that signal. Under independent on-policy sampling, we derive the exact objective followed by BLOO's mean update, making the prompt weighting induced by predictive scaling explicit and allowing us to quantify the contribution of homogeneous groups. Their share of the per-prompt mean coefficient is bounded sharply by $1/(2K)$ and reaches only $7.47\%$ in the example above, despite their high frequency, while a full gradient-variance decomposition shows how differences among equally rewarded responses can add noise to the retained update. On Qwen3-VL-4B trained on ViRL39K, BLOO improves final validation accuracy over GRPO by $0.56$ and $1.47$ percentage points in the with-reference and KL-disabled configurations, respectively. The corresponding gains averaged over the final $24$ training steps are $1.98$ and $0.57$ percentage points, providing evidence that the observed advantage extends beyond the final checkpoint.

## 2 Problem Setup and Motivation

To explain how group rewards become learning signals, we distinguish the reward of an individual response from the policy's success probability and the parameter update estimated from a sampled group. For a fixed prompt $x$, draw $K \geq 2$ independent responses $Y_i$ from the current policy $\pi_\theta$ and evaluate them with a fixed binary verifier $R_i = r_x(Y_i) \in \{0, 1\}$. Writing $S = \sum_i R_i$ and $p = p_\theta(x)$, the observed success fraction $S/K$ varies across groups, whereas $p$ remains fixed for $x$ and policy.

**From a reward to an update.** The response's *score* $Z_i$ denotes the gradient of its log probability, so weighting it by a positive scalar credit $A_i$ encourages that response, while a negative credit discourages it. We distinguish the policy's success probability from its sampled estimate $S/K$ and define the response score $Z_i$, true reward gradient $g_x$, and credit-weighted group update $\widehat{g}_x$ by

$$\begin{aligned} p_\theta(x) = \mathbb{E}[R_i \mid x], \quad Z_i = \nabla_\theta \log \pi_\theta(Y_i \mid x), \\ g_x = \nabla_\theta p_\theta(x) = \mathbb{E}[R_i Z_i \mid x], \quad \widehat{g}_x = \frac{1}{K} \sum\nolimits_{i=1}^{K} A_i Z_i. \end{aligned} \tag{1}$$

Here $g_x$ is the true reward gradient and $\widehat{g}_x$ is the credit-weighted estimate from a group, its mean and variation over repeated sampling determine the signal and noise analyzed below. We assume finite score second moments and treat credits as detached, *i.e.*, as constants when differentiating the actor loss, with the regularity conditions for exchanging differentiation and expectation stated in the § B.1.

**Where empirical contrast disappears.** With full-group empirical mean $m^{\mathrm{G}} = S/K$ and variance $(\sigma^{\mathrm{G}})^2 = m^{\mathrm{G}}(1 - m^{\mathrm{G}})$, the binary-reward credits assignment of GRPO Shao et al. (2024) and REINFORCE Leave-One-Out (RLOO) (Ahmadian et al., 2024) are

$$A_i^{\mathrm{G}} = \begin{cases} (R_i - m^{\mathrm{G}})/\sigma^{\mathrm{G}}, & 0 < S < K, \\ 0, & S \in \{0, K\}, \end{cases} \qquad A_i^{\mathrm{R}} = R_i - \frac{S - R_i}{K - 1}. \tag{2}$$

Both rules give zero credit on the *homogeneous* event $H = \{S = 0\} \cup \{S = K\}$ with probability $q_K(p) = p^K + (1 - p)^K$, so comparison signals can frequently disappear even when $g_x \neq 0$.

To address this gap, we construct the reference and scale without the target, using independence to simplify the mean update and avoid the $1 - 1/K$ attenuation caused by unscaled full-group centering, as shown in Appendix B. We combine this exclusion with predictive smoothing to reflect the limited sibling evidence, so an all-failure sample does not force the estimated success probability to zero.

## 3 BLOO: A Posterior Predictive Construction

To satisfy two requirements in §2, we establish how sibling-only statistics preserve the mean update direction, then use a Bayesian prediction to choose the reference and scale from finite sibling evidence.

### 3.1 A sibling reference preserves the average direction

To determine how a reference affects the expected update, let $B_{-i}$ be a baseline and $C_{-i}$ a positive inverse scale computed from siblings, whose independence from the target makes the baseline contribution vanish and yields the following mean identity.

**Theorem 1** (Sibling-only mean identity). *For sibling-only $B_{-i}$ and $C_{-i} > 0$, the sampling and regularity conditions above together with the integrability conditions in Appendix B.1 give*

$$\mathbb{E}[C_{-i}(R_i - B_{-i})Z_i \mid x] = \mathbb{E}[C_{-i} \mid x]\, g_x. \tag{3}$$

At fixed $x, \theta$, the target is independent of its siblings and $\mathbb{E}[Z_i \mid x] = 0$, so the subtracted baseline contributes zero to the expected update, leaving the reward gradient multiplied by the positive average inverse scale. The identity therefore holds even when the reference does not estimate $p$ perfectly, although its accuracy still affects variance and any individual group can produce an unfavorable update. The identity permits a predictive reference with an explicit mean direction, a property also shared by GRPO and RLOO methods, whose key differences from BLOO are summarized in Table 5.

### 3.2 Finite evidence defines the prediction and its scale

In small groups, an empirical average assigns zero probability to unobserved outcomes whenever all sampled peers succeed or fail. To keep unobserved outcomes possible, a working $\mathrm{Beta}(1/2, 1/2)$ model adds half a success and half a failure to the $K - 1$ sibling counts (Jeffreys, 1946), yielding a success prediction $m_{-i}$ strictly between zero and one. BLOO divides the target's prediction error by the next reward's standard deviation $\sigma_{-i} = \sqrt{m_{-i}(1 - m_{-i})}$, giving

$$m_{-i} = \frac{S - R_i + 1/2}{K}, \qquad A_i^{\mathrm{B}} = \frac{R_i - m_{-i}}{\sigma_{-i}}. \tag{4}$$

Where the numerator measures the target reward's deviation from the sibling prediction, while the denominator captures the intrinsic predictive uncertainty of the next binary outcome rather than the posterior uncertainty of $U$, which is smaller by a factor of $K + 1$. Theorem 1 guarantees that the expected update direction remains aligned with the true reward gradient, regardless of whether the prior is well-specified. Beyond stabilizing the mean direction, predictive smoothing keeps $m_{-i} \in (0, 1)$, preventing zero denominators when all rewards agree.

**Proposition 1** (Finite credit from finite evidence). *For every binary group with $K \geq 2$,*

$$|A_i^{\mathrm{B}}| \leq \sqrt{2K - 1}, \qquad A_i^{\mathrm{B}} = \begin{cases} -1/\sqrt{2K - 1}, & S = 0, \\ +1/\sqrt{2K - 1}, & S = K. \end{cases} \tag{5}$$

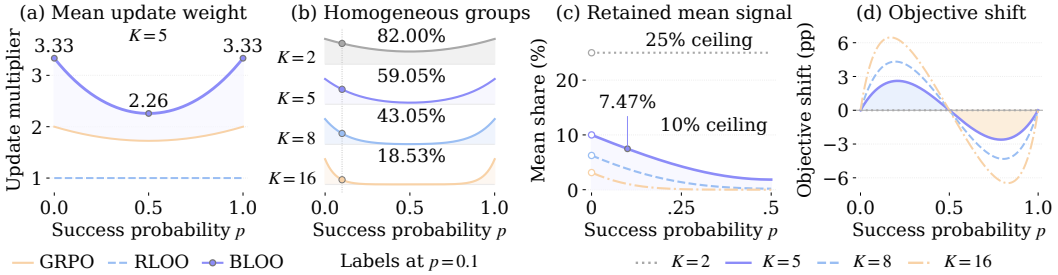


Figure 2: (a) BLOO gives larger mean-update multipliers near $p = 0$ or 1, while (b,c) show that homogeneous groups can be frequent yet contribute little mean signal, with $59.05\%$ of groups supplying only $7.47\%$ of the mean coefficient at $K = 5, p = 0.1$ (ceiling $10\%$). (d) The normalized objective departs from the mean reward, reflecting prompt reweighting by predictive scaling.

This bound arises because smoothing constrains $m_{-i}$ strictly within $[1/(2K), 1 - 1/(2K)]$. As illustrated in Figure 1(b)(c), while agreement yields smaller updates than unexpected outcomes, we note that bounding $A_i^{\mathrm{B}}$ controls the scalar credit rather than the full update $A_i^{\mathrm{B}} Z_i$. In practice, BLOO only requires computing the group sum $S$ once, and detaching credits to weight each response's sum of token-score gradients, which operates in $O(K)$ time with zero additional rollouts or historical caches. The policy term equals Equation 1 at rollout parameters, whereas length averaging, clipping, KL regularization, and policy reuse can alter the analyzed update, as detailed in Appendix C.

## 4 Signal Retention: Mean Objective and Noise

To assess retained credit beyond its availability, we identify the mean objective, quantify homogeneous groups' contribution, and separate two sources of gradient noise, with full proofs in the Appendix B.

### 4.1 Predictive Scaling Determines the Prompt Objective

Although Theorem 1 preserves each prompt's mean direction, its multiplier can vary across prompts and thus change the aggregate objective when they share a policy. Building on the interpretation of normalization as adaptive gradient scaling (Ge et al., 2026), we compute BLOO's weighting exactly by averaging over possible sibling success counts. Specifically, let $b_{n,s}(p) = \binom{n}{s} p^s (1-p)^{n-s}$ denote the probability of $s$ successes in $n$ trials and $m_s = (s + 1/2)/K$ the corresponding sibling prediction, so the expectation in Theorem 1 becomes the finite sum below.

**Theorem 2** (Expected update and transformed objective)**.** *For BLOO,*

$$\mathbb{E}[\widehat{g}_x^{\mathrm{B}} \mid x] = \alpha_K(p)\, g_x,$$
$$\alpha_K(p) = \sum_{s=0}^{K-1} \frac{b_{K-1,s}(p)}{\sqrt{m_s(1-m_s)}}, \quad 2 \le \alpha_K(p) \le \frac{2K}{\sqrt{2K-1}}. \tag{6}$$

*Define* $\Psi_K(p) = \int_0^p \alpha_K(u)\, du$*. For a fixed prompt distribution with the regularity in Appendix B.1,*

$$F_K(\theta) = \mathbb{E}_x[\Psi_K(p_\theta(x))], \qquad \mathbb{E}_x \mathbb{E}[\widehat{g}_x^{\mathrm{B}} \mid x] = \nabla_\theta F_K(\theta). \tag{7}$$

Where each term weights a sibling count's inverse scale by its probability, making $\alpha_K(p)$ the expected update strength relative to the true reward gradient. Substituting BLOO's inverse scale into Theorem 1 gives Equation 6, and defining $\Psi_K$ to have slope $\alpha_K(p)$ and applying the chain rule then yields Equation 7, thus connecting the local weighting to the objective followed by the expected update.

Since the mean multiplier changes the update strength and the resulting prompt objective, as shown in Figure 2(a)(d), Section 5.6 matches the estimators' mean strength to compare noise without confounding it with scaling. Moreover, prompts can favor conflicting directions in shared parameters, meaning that an increase in $F_K$ need not increase raw mean reward, as the two-prompt counterexample in Appendix B.6 demonstrates. Related analyses examine pass@$k$ prompt reweighting (Barakat et al., 2026) and uncertainty-aware gradient aggregation (Guo et al., 2026), while our theorem identifies BLOO's particular objective without assuming that its weighting resolves such conflicts.

### 4.2 Frequent homogeneous groups supply a bounded mean contribution

To identify which groups drive the mean update, we distinguish the probability $q_K(p)$ of a homogeneous group from its contribution after averaging. Specifically, we measure this contribution via the unconditional expectation of the update restricted to $H$, zeroing outcomes outside $H$.

**Proposition 2** (Recovered homogeneous contribution). *For $H = \{S = 0\} \cup \{S = K\}$,*

$$\mathbb{E}[\mathbf{1}_H \widehat{g}_x^{\mathrm{B}} \mid x] = h_K(p)\, g_x, \quad h_K(p) = \frac{p^{K-1} + (1-p)^{K-1}}{\sqrt{2K-1}}, \quad 0 < \frac{h_K(p)}{\alpha_K(p)} \le \frac{1}{2K}. \tag{8}$$

An all-failure group discourages the sampled failures and an all-success group reinforces the sampled successes, yielding a positive multiple of $g_x$ when averaged over repeated groups. The retained credit therefore contributes mean signal without revealing an unseen correct answer or guaranteeing improvement from any individual homogeneous group. Because the total mean and homogeneous contribution both point along $g_x$, their scalar coefficients can be compared directly, as shown in Figure 2 (b)(c). At $K = 5, p = 0.1$, homogeneous groups occur $59.05\%$ of the time but supply only $7.47\%$ of BLOO's mean coefficient, with their share bounded by $10\%$ at this group size. The general bound follows because the two endpoint terms in $\alpha_K$ sum to $2Kh_K$ and all remaining terms are nonnegative; Appendix B.7 establishes sharpness and explains why both mean contributions vanish when $g_x = 0$, despite their positive coefficients.

### 4.3 Equal rewards can still produce different update directions

To assess the noise accompanying the mean signal, we distinguish variation in the number of successes from variation in score gradients among responses with the same reward. The latter is especially relevant to homogeneous groups, because two failed solutions receive the same scalar credit but can move the parameters in different directions when discouraged. Let $a_r(s)$ denote the credit for reward $r$ in a group with $s$ successes and $\Sigma_r = \mathrm{Cov}(Z_i \mid R_i = r, x)$ the score covariance within that reward class, which captures variation invisible to the binary labels.

To separate these sources, first condition on the success count, giving $\mathbb{E}[\widehat{g}_x \mid S = s, x] = u(s; p) g_x$ for $0 < p < 1$, where $u(s; p) = s\, a_1(s)/(Kp) - (K - s)a_0(s)/(K(1-p))$ combines successful and failed responses, with terms for absent outcomes omitted. Applying the law of total variance then yields the following decomposition of the full group update.

**Theorem 3** (Two sources of group noise). *For $0 < p < 1$, any count-based credit rule with finite second moments, and $S \sim \mathrm{Binomial}(K, p)$,*

$$\mathrm{tr}\,\mathrm{Cov}(\widehat{g}_x \mid x) = \underbrace{\mathrm{Var}_S[u(S; p)]\, \|g_x\|_2^2}_{\textit{variation in success count}} + \underbrace{\frac{1}{K^2}\mathbb{E}_S\left[S\, a_1(S)^2\, \mathrm{tr}\,\Sigma_1 + (K - S)a_0(S)^2\, \mathrm{tr}\,\Sigma_0\right]}_{\textit{variation among responses with the same reward}}. \tag{9}$$

The left side measures the expected squared distance from the mean update, with the first term capturing changes in the success count and the second capturing the remaining variation at a fixed count. The squared-credit factors in the second term explain why a bounded, nonzero credit can still transmit substantial score variation. As the decomposition accounts for dependence induced by shared sibling statistics, it describes the full group gradient. The homogeneous credits that preserve the mean contribution in Proposition 2 also retain variation among equally rewarded responses, exposing a trade-off between recovered signal and additional noise. This noise source vanishes when their score gradients are identical but can persist with diverse failures even as success becomes rare and the reward gradient shrinks. Appendix B.8 proves a regime in which BLOO is noisier than GRPO and RLOO, thus showing why finite credit and a correct mean direction do not guarantee lower variance.

## 5 Experiments

### 5.1 Experimental Settings

We train Qwen3-VL-4B (Bai et al., 2025) on Geometry3k[1] and ViRL39K (Wang et al., 2026a). Training uses one epoch, a learning rate of $5 \times 10^{-5}$, 8 H20 GPUs, and the EasyR1(Zheng et al., 2025c)

[1] https://huggingface.co/datasets/hiyouga/geometry3k

Table 1: Qwen3-VL-4B accuracy (%) on seven benchmarks. Best / second-best scores are bold/underlined. Math/Gen Avg. average five/two tasks; Avg. averages all seven.

| Method | Visual Mathematical Reasoning | | | | | | General Reasoning | | | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geo3K | MathVista | MathVision | MathVerse | We-Math | **Math Avg.** | MMMU | MMStar | **Gen Avg.** | |
| Qwen3-VL-4B | 64.4 | 77.0 | 53.2 | 63.6 | 78.6 | 67.4 | 63.1 | 69.7 | 66.4 | 67.1 |
| SFT | 62.4 (.4) | 76.5 (.1) | 52.6 (.2) | 63.0 (.2) | 78.1 (.2) | 66.5 | 62.3 (.3) | 69.1 (.2) | 65.7 | 66.3 |
| GRPO | 65.5 (.8) | 77.4 (.3) | 53.4 (.3) | 64.0 (.4) | 78.8 (.3) | 67.8 | 63.2 (.4) | 69.9 (.3) | 66.6 | 67.5 |
| RLOO | 66.0 (.8) | 77.5 (.3) | 53.7 (.3) | 64.3 (.4) | 79.0 (.3) | 68.1 | 63.1 (.4) | 70.1 (.3) | 66.6 | 67.7 |
| Dr.GRPO | 65.3 (.9) | 77.5 (.3) | 53.8 (.4) | 64.1 (.4) | 79.1 (.3) | 68.0 | 63.4 (.5) | 70.1 (.4) | 66.8 | 67.6 |
| DBB | 67.1 (.6) | 78.0 (.3) | 54.2 (.3) | 65.1 (.3) | 79.6 (.2) | 68.8 | 64.0 (.4) | 70.7 (.3) | 67.4 | 68.4 |
| BLOO | **68.7** (.5) | **78.4** (.2) | **54.8** (.3) | **65.8** (.2) | **80.1** (.2) | **69.6** | **64.5** (.4) | **71.2** (.3) | **67.9** | **69.1** |

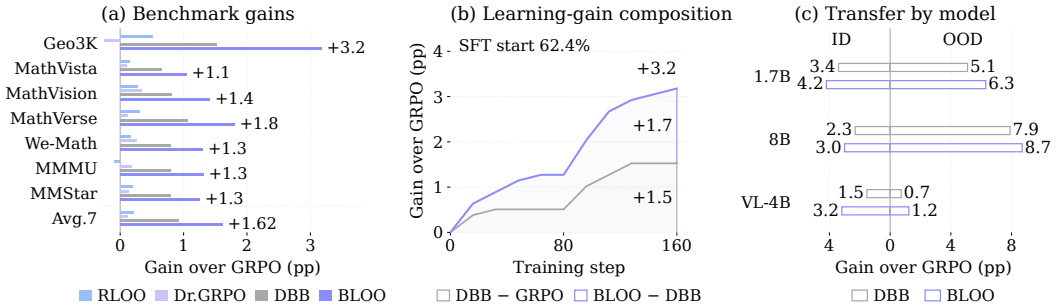


Figure 3: Comprehensive benchmark and transfer comparisons. (a) Gains over GRPO; (b) Geometry3K gains split into DBB–GRPO and BLOO–DBB; (c) ID/OOD gains shown left/right, with visual OOD averaging MathVista and MathVision. Gains use unrounded scores and are in pp.

framework. We evaluate on ten benchmarks covering mathematical reasoning and general multimodal understanding. WeMath, MathVista, MathVerse, MathVision assess mathematical reasoning, MMMU and MMStar assess broader visual understanding and reasoning. Table 1 reports overall performance.

We compare BLOO with GRPO, RLOO, Dr.GRPO, and DBB, which respectively use group normalization, empirical sibling centering, unscaled centering, and historical reward evidence. The matched visual comparison uses $K = 5$ with shared initialization, rollout budgets, and optimization settings. Appendix D.1 gives their provenance and the evaluation protocol.

### 5.2 Benchmark Comparisons

In the main comparison in Table 1, BLOO ranks first on all seven benchmarks, with an average of 69.1%, compared with 67.5% for GRPO and 68.4% for DBB. Its mathematical and general reasoning averages reach 69.6% and 67.9%, respectively, so the improvement spans both benchmark groups. The original-model and SFT averages of 67.1% and 66.3% provide additional reference points for assessing net improvement after post-training. Figure 3 (a) shows the largest gain over GRPO on Geometry3K (3.2 percentage points, pp), while subfigure (c) gives a 1.2 pp gain on the MathVista/MathVision transfer average. Subfigure (b) separates the gains of DBB over GRPO and BLOO over DBB throughout training. Together, these comparisons test whether better credit assignment improves the training-domain task and carries over to other visual reasoning tasks.

### 5.3 Observed Training on ViRL39K

Figure 4 compares BLOO and GRPO in the supplied ViRL39K runs. At step 120, BLOO reaches 72.32% validation accuracy versus 71.76% with reference, and 75.15% versus 73.68% with KL disabled. Its average validation advantage over the final 24 steps is 1.98 and 0.57 pp, respectively. BLOO therefore performs better at both the endpoint and over the final training window in each recorded configuration, providing initial evidence of an accuracy advantage beyond the final checkpoint.

Subfigures (b)(d) show training-batch correctness alongside validation in subfigures (a)(c). Endpoint and window averages use raw values; smoothing is only for display. Each configuration contains one

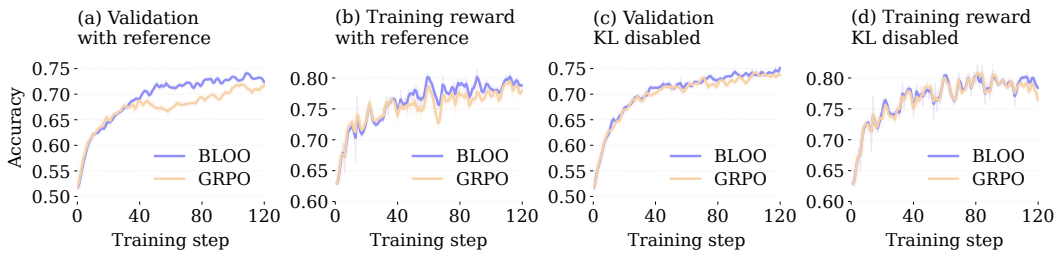


Figure 4: Observed Qwen3-VL-4B trajectories on ViRL39K. Validation and training-batch accuracy with reference (a,b) and KL disabled (c,d). Faint lines are raw values; thicker lines apply Gaussian smoothing ($\sigma = 0.8$ steps). Each curve is one run, with accuracy shown as a fraction.

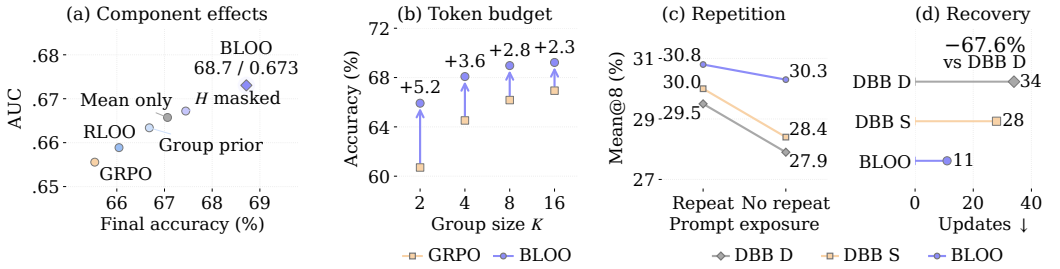


Figure 5: Different components and stress-test comparisons. (a) Final accuracy versus AUC; (b) fixed-budget gains over GRPO; (c) sensitivity to prompt repetition; (d) recovery after a curriculum shift. Panels (c,d) use Qwen3-1.7B; D/S denotes default/searched DBB discounts.

run per method, so the results support these within-configuration comparisons without establishing multi-seed significance or isolating the effect of KL.

### 5.4 Component Analysis

Table 2 tests the roles of the sibling reference, prior smoothing, predictive scaling, and homogeneous-group credit under a common token-sum objective. RLOO removes the target from an empirical baseline, full-group Jeffreys adds smoothing while retaining the target, and mean-only BLOO removes predictive scaling. The masked variant sets only homogeneous-group credits to zero. In the comparison, full BLOO reaches 68.7%, versus 67.1% without scaling and 67.4% with homogeneous groups masked. These comparisons isolate the contributions of scaling and retained credit beyond target exclusion alone.

Table 2: Ablations study on Qwen3-VL-4B model. AUC averages accuracy over training, drop is peak minus final accuracy.

| Variant | Accuracy (%) | AUC | Peak-final drop |
|---|---|---|---|
| GRPO | 65.5 | .656 | 1.5 |
| Empirical sibling mean (RLOO) | 66.0 | .659 | 1.3 |
| Full-group Jeffreys | 66.7 | .663 | 1.0 |
| BLOO mean only | 67.1 | .666 | 0.9 |
| BLOO masking $S \in \{0, K\}$ | 67.4 | .667 | 0.8 |
| Full BLOO | **68.7** | **.673** | **0.4** |

In Figure 5(a), we plot final accuracy against the area under the learning curve (AUC) to assess sustained performance across the entire training trajectory, rather than relying on a single final checkpoint. Full BLOO achieves the highest AUC (0.673), outperforming both mean-only BLOO (0.666) and the homogeneous-masked variant (0.667). This gap indicates that retaining credit provides a continuous advantage throughout training. Crucially, because predictive scaling alters the magnitude of gradient updates, matched learning-rate tuning is necessary to disentangle the benefits of the scaling mechanism from simply taking larger optimization steps. Appendix Table 8 further reports analytic token-utilization rates and gradient norms to support this analysis.

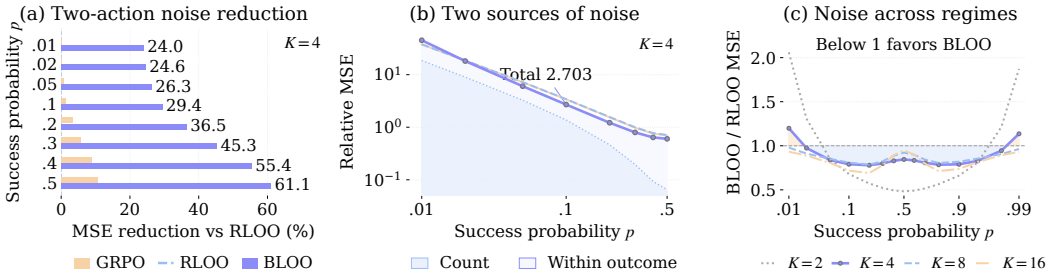


Figure 6: Exact gradient-noise comparisons after matching mean updates. (a) Two-action MSE reduction versus RLOO; (b) four-action count noise and total noise; (c) BLOO/RLOO MSE ratios across success probabilities.

### 5.5 Group Size and Training Stability

With a fixed generated-token budget, larger groups provide more sibling evidence but cover fewer prompts. Table 3 and Figure 5(b) show a BLOO–GRPO gap of 5.2 pp at $K = 2$ and 2.3 pp at $K = 16$. The observed advantage is largest when sibling evidence is limited, which directly matches BLOO's motivation. Increasing $K$ from 8 to 16 adds only about 0.3 pp to BLOO's score while raising its analytic credit bound from 3.87 to 5.57. The overhead column estimates extra scalar-credit computation; it does not report measured end-to-end runtime.

Table 3: Group-size comparison on Qwen3-VL-4B with a fixed generated-token budget.

| $K$ | GRPO Acc. | BLOO Acc. | Gain | BLOO $\lvert A\rvert$ bound | Overhead |
|---|---|---|---|---|---|
| 2 | 60.7 | **65.9** | +5.2 | 1.73 | 0.16% |
| 4 | 64.5 | **68.1** | +3.6 | 2.65 | 0.18% |
| 8 | 66.2 | **69.0** | +2.8 | 3.87 | 0.21% |
| 16 | 66.9 | **69.2** | +2.3 | 5.57 | 0.27% |

Table 4 checks whether an advantage persists late in training and after extending the budget. BLOO's final and last-20% accuracies are close (68.7% and 68.8%), with a peak-to-final drop of 0.4 pp versus 1.5 pp for GRPO. After twice the training horizon, the accuracies are 68.5% and 64.1%, respectively. This comparison tests sustained accuracy, while the gradient-norm percentile tracks large updates separately.

Table 4: Qwen3-VL-4B stability on Geometry3K. Last-20% averages the final training fifth; drop is peak minus final accuracy in pp; Grad P99 is the 99th-percentile gradient norm; 2x final uses twice the training horizon.

| Method | Final | Last-20% | Drop ↓ | Grad P99 ↓ | 2x final |
|---|---|---|---|---|---|
| GRPO | 65.5 (.8) | 65.8 | 1.5 | .29 | 64.1 |
| DBB | 67.1 (.6) | 67.3 | 0.8 | .25 | 66.4 |
| BLOO | **68.7 (.5)** | **68.8** | **0.4** | **.22** | **68.5** |

Figure 5 (c) and (d) further test reliance on historical evidence. In the appendix Qwen3-1.7B comparison, removing prompt repetition reduces BLOO's accuracy by 0.5 pp versus 1.6 pp for tuned DBB, and recovery after a curriculum shift takes 11 versus 28 updates. These tests assess whether using current sibling evidence helps when past rewards become less informative; the full comparison appears in Appendix Table 9.

### 5.6 Exact Gradient Diagnostics

To explain when retained credit helps, we enumerate two- and four-action tasks with known reward gradients. The four-action task allows responses with the same reward to have different score gradients. Checks over 2,044 binary groups and 120 task settings recover the mean identities to within $10^{-12}$ (Appendix C). We then rescale each estimator to the same mean gradient and compare its relative mean-squared error (MSE), so the comparison measures noise independently of update magnitude. This oracle rescaling is used only for diagnosis.

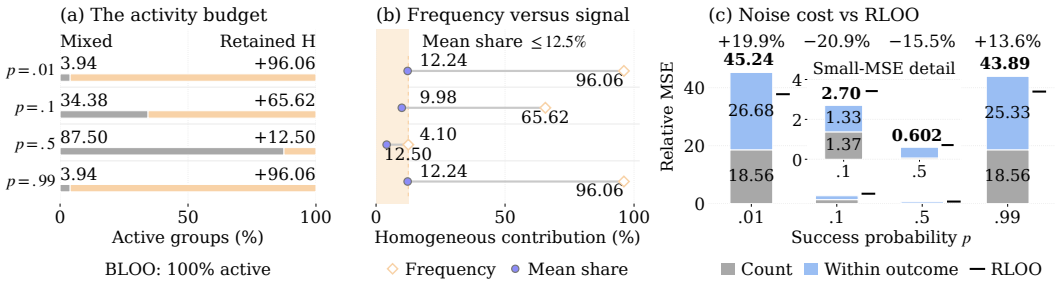


Figure 7: Credit availability, retained signal, and gradient noise at $K = 4$. (a) Proportion of active groups yielding non-zero credit, partitioned into mixed and retained homogeneous cases; (b) comparison between homogeneous-group frequency (open markers) and its retained share in the expected update (filled markers), bounded by the theoretical $12.5\%$ ceiling $(1/(2K))$; (c) gradient MSE decomposed into count and within-outcome noise, with RLOO baseline ticks for comparison.

Figure 6(a) shows lower MSE for BLOO in the two-action task. With within-reward gradient variation, however, the benefit depends on the policy's success probability. At $K = 4$, BLOO reduces MSE relative to RLOO by $20.95\%$ at $p = 0.1$ but increases it by $13.59\%$ at $p = 0.99$, as shown in panels (b,c). Retaining credit can therefore improve gradient quality, but it does not guarantee lower variance.

Figure 7 connects this trade-off to homogeneous groups. At $K = 4, p = 0.1$, BLOO increases the fraction of groups with nonzero credit from $34.38\%$ to $100\%$, while homogeneous groups contribute only $9.98\%$ of its mean coefficient, below the $12.5\%$ bound. Subfigure (c) shows the accompanying noise cost. BLOO thus uses feedback that empirical centering discards while keeping its mean contribution bounded, and the value of that feedback still depends on the response gradients.

## 6 Related Work

Group-based RLVR assigns credit using responses to the same prompt, with GRPO normalizing group rewards and RLOO using a sibling baseline (Shao et al., 2024; Ahmadian et al., 2024). Dr.GRPO and DAPO further modify normalization and sampling (Liu et al., 2025; Yu et al., 2025), while Bayesian approaches use Beta-based normalization, global shrinkage, or discounted reward history (Xiao et al., 2025; Han et al., 2026; Kim et al., 2026). Building on these ideas, BLOO derives both a smoothed prediction and the next-reward scale from current siblings, retaining finite credit in homogeneous groups without additional rollouts or stored history. Its analysis identifies the resulting mean objective and quantifies the retained signal and its variance cost. Appendix F provides the full comparison.

## 7 Conclusion

In group-relative policy optimization, relying on raw empirical baselines inevitably causes gradient starvation whenever sampled responses produce identical binary outcomes. To overcome this limitation, BLOO utilizes a Jeffreys-smoothed prediction computed solely from sibling responses, scaling each target's error by the predictive standard deviation of the next binary outcome. This formulation ensures finite, bounded credit even in all-failure or all-success groups without needing extra rollouts or learned value networks. Beyond guaranteeing active gradients, our theoretical framework reveals what this retained signal actually does: we derive the explicit mean objective induced by predictive scaling, prove that homogeneous groups contribute at most a bounded fraction to the expected update direction, and isolate the hidden noise introduced by diverse responses sharing identical binary labels.

## Limitations

BLOO focuses on binary verifier rewards, with its exact analysis assuming independently sampled on-policy responses. Extending the predictive credit construction to continuous rewards or process-level feedback is a natural direction for future work. The current empirical scope centers on visual reasoning, and broader evaluation across additional model families and tasks would further clarify where the method is most effective. (*We provide the AI Use Statement in Section G*)

## REFERENCES


Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. In *Proceedings of ACL*, 2024.

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

Anas Barakat, Souradip Chakraborty, Khushbu Pahwa, and Amrit Singh Bedi. Why Pass@k Optimization Can Degrade Pass@1: Prompt Interference in LLM Post-training. *arXiv preprint arXiv:2602.21189*, 2026. URL https://arxiv.org/abs/2602.21189.

Michael Bereket and Jure Leskovec. Uncalibrated Reasoning: GRPO Induces Overconfidence for Stochastic Outcomes. *arXiv preprint arXiv:2508.11800*, 2025. URL https://arxiv.org/abs/2508.11800.

Lin Chen, Jinsong Li, Xiaoyi Dong, et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2403.20330.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Zeyun Deng, Yuzhe Lu, Yawei Wang, Linbo Liu, Qing Ping, Han Ding, Guande Wu, Panpan Xu, and Jun Huan. Prism-GRPO: Faster VLA Policy Optimization via Splitting Same-outcome Groups. *arXiv preprint arXiv:2608.17423*, 2026. URL https://arxiv.org/abs/2608.17423.

Cheng Ge, Caitlyn Heqi Yin, Hao Liang, and Jiawei Zhang. Why GRPO Needs Normalization: A Local-Curvature Perspective on Adaptive Gradients. *arXiv preprint arXiv:2601.23135*, 2026. URL https://arxiv.org/abs/2601.23135.

Peizheng Guo, Jianqi Zhang, Xingyu Zhang, Yun Fan, Jiahuan Zhou, Changwen Zheng, and Wenwen Qiang. Gupo: Gradient uncertainty-aware policy optimization for post-training large language models. *arXiv preprint arXiv:2608.17411*, 2026.

Kevin Han, Yuhang Zhou, Mingze Gao, Gedi Zhou, Serena Li, Abhishek Kumar, Xiangjun Fan, Weiwei Li, and Lizhu Zhang. EBPO: Empirical Bayes Shrinkage for Stabilizing Group-Relative Policy Optimization. *arXiv preprint arXiv:2602.05165*, 2026. URL https://arxiv.org/abs/2602.05165.

Chaoqun He, Renjie Luo, Yuzhuo Bai, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. *arXiv preprint arXiv:2605.21125*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring mathematical problem solving with the math dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A*, 186(1007):453–461, 1946.

Haechan Kim, Soohyun Ryu, Gyouk Chu, Doohyuk Jang, and Eunho Yang. Discounted beta–bernoulli reward estimation for sample-efficient reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2603.18444*, 2026.

Yu Li, Rui Miao, Tian Lan, and Zhengling Qi. OPPO: Bayesian Value Recursion for Token-Level Credit Assignment in LLM Reasoning. *arXiv preprint arXiv:2605.21851*, 2026. URL https://arxiv.org/abs/2605.21851.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of ACL-IJCNLP*, pp. 6774–6786, 2021.

Pan Lu, Hritik Bansal, Tony Xia, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Runqi Qiao, Qiuna Tan, Guanting Dong, et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of ACL*, 2025. URL https://arxiv.org/abs/2407.01284.

David Rein, Betty Li Hou, Asa Cooper Stickland, et al. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

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

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

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Shota Takashiro, Soichiro Nishimori, Paavo Parmas, et al. On advantage estimates for max@$k$ policy gradients. *arXiv preprint arXiv:2606.06080*, 2026.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning. *arXiv preprint arXiv:2504.08837*, 2025. URL https://arxiv.org/abs/2504.08837.

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

Jiamian Wang, Samyadeep Basu, Koustava Goswami, Tong Yu, and Zhiqiang Tao. Spurious advantage hidden in grpo. *arXiv preprint arXiv:2609.04063*, 2026b.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*, 2024a.

Yubo Wang, Xueguang Ma, Ge Zhang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. BNPO: Beta Normalization Policy Optimization. *arXiv preprint arXiv:2506.02864*, 2025. URL https://arxiv.org/abs/2506.02864.

Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of CVPR*, 2024. URL https://arxiv.org/abs/2311.16502.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Zhi Zhang, Zhen Han, Costas Mavromatis, Qi Zhu, Yunyi Zhang, Sheng Guan, Dingmin Wang, Xiong Zhou, Shuai Wang, Soji Adeshina, et al. Train less, learn more: Adaptive efficient rollout optimization for group-based reinforcement learning. *arXiv preprint arXiv:2602.14338*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Haizhong Zheng, Yang Zhou, Brian R. Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. *arXiv preprint arXiv:2506.02177*, 2025b.

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

# A Notation and Reading Guide

Using the notation below, the appendix follows the main text from the estimator definition and predictive construction to the mean contribution and variance, providing the derivations needed to connect each stage.

| Symbol | Meaning |
|---|---|
| $x, \mathcal{D}$ | Prompt and fixed prompt distribution. |
| $\theta, \theta_{\text{old}}$ | Current parameters and rollout-policy parameters. |
| $Y_i, R_i, T_i$ | Response trajectory, binary verifier reward, valid response length. |
| $K, S, S_{-i}$ | Group size, total success count, sibling success count. |
| $R_{-i}$ | Sibling reward vector $(R_j)_{j\neq i}$. |
| $p_\theta(x), p$ | Fixed-policy success probability; $p$ abbreviates it locally. |
| $U, \widetilde{R}_i$ | Working Bayesian latent probability and predictive reward. |
| $Z_i, g_x$ | Sequence score and true per-prompt reward gradient. |
| $\widehat{g}_x$ | Realized group-average update $K^{-1}\sum_i A_i Z_i$. |
| $B_{-i}, C_{-i}$ | Generic sibling baseline and positive inverse scale. |
| $m^{\text{G}}, \sigma^{\text{G}}$ | Full-group empirical mean and standard deviation (GRPO). |
| $m_{-i}, \sigma_{-i}$ | Sibling posterior mean and predictive standard deviation. |
| $A_i^{\text{B}}, A_i^{\text{G}}, A_i^{\text{R}}$ | BLOO, GRPO, and RLOO sequence credits. |
| $b_{n,s}(p), m_s, c_s$ | Binomial mass, count-indexed mean, and inverse scale. |
| $\alpha_K, \gamma_K, h_K$ | BLOO, GRPO, and homogeneous-event gradient multipliers. |
| $\Psi_K, F_K$ | Prompt transformation and aggregate transformed objective. |
| $H, q_K(p)$ | Homogeneous event and its probability. |
| $a_r(s), u(s;p), \Sigma_r$ | Count-based credit, conditional mean multiplier, within-outcome score covariance. |
| $\mathbb{E}_{\text{old}}$ | Prompt and iid rollout expectation under the fixed old policy. |
| $\rho_{i,t}, \epsilon_\ell, \epsilon_h, \beta$ | Token ratio, lower/upper clip widths, KL coefficient. |

**A route through the proofs.** Appendix B.1 establishes the sampling conditions and score identities behind Theorem 1, after which Appendices B.4 and B.5 construct the prediction, explain its scale, and derive finite credit bounds. Appendices B.6–B.8 then identify the mean objective, quantify the homogeneous contribution, and explain its noise cost, with Appendix C connecting the results to the actor surrogate and numerical checks.

**Two probability models, two roles.** The Bayesian model uses a latent probability $U$ to construct a prediction from finite observations, whereas the sampling model fixes $x, \theta$ and studies repeated groups from the actual policy with success probability $p_\theta(x)$. Predictive centering holds under the former and the gradient identities under the latter, so the gradient analysis does not require the policy's true success probability to follow the chosen prior distribution.

# B Complete Theoretical Derivations

## B.1 Assumptions and score identities

Starting from Equation 1, write $p = p_\theta(x)$, $S = \sum_i R_i$, $S_{-i} = S - R_i$, and $\widehat{g}_x = K^{-1}\sum_i A_i Z_i$, with the prompt and policy parameters held fixed in all per-prompt expectations.

Fix an open parameter set $\Theta \subset \mathbb{R}^d$ and a conditional law $\pi_\theta(\cdot \mid x)$ with parameter-independent support that is differentiable almost everywhere, together with a fixed binary verifier $r_x$ having no direct parameter dependence. Assume that trajectories are conditionally iid, $\mathbb{E}[\|Z_i\|_2^2 \mid x] < \infty$, and differentiation commutes locally with integration for normalization and expected reward. For prompt averaging, require $\mathcal{D}$ to be independent of $\theta$ and an integrable local envelope to permit exchanging prompt expectation and gradient, while the generic sibling rule satisfies

$$\mathbb{E}[C_{-i} \mid x] < \infty, \qquad \mathbb{E}[|C_{-i}B_{-i}| \mid x] < \infty, \qquad C_{-i} > 0 \text{ a.s.} \tag{10}$$

These conditions justify the score identities below, which give a zero expected score and identify the expected reward-weighted score with the reward gradient.

$$\mathbb{E}[Z_i \mid x] = \nabla_\theta \int \pi_\theta(dy \mid x) = 0, \qquad \mathbb{E}[R_i Z_i \mid x] = \nabla_\theta \int r_x(y)\pi_\theta(dy \mid x) = g_x. \tag{11}$$

Because the identities concern independent samples from the stated policy, changes induced by adaptive acceptance, diverse beam search, or shared decoding randomness require separate analysis, and filtering groups by their rewards does not automatically preserve the assumptions.

### B.2 Proof of the attenuation identity and Theorem 1

To isolate the effect of including the target in its own baseline, we subtract the full-group mean before applying Equation 2's population-divisor scale and homogeneous zero-credit convention, obtaining the attenuation in Equation 12.

$$\mathbb{E}[(R_i - m^{\mathrm{G}})Z_i \mid x] = \left(1 - \frac{1}{K}\right) g_x. \tag{12}$$

Let $\mathcal{F}_{-i} = \sigma(Y_j : j \neq i)$ denote the information supplied by the other responses, so a sibling-measurable statistic depends only on $\mathcal{F}_{-i}$ and remains detached during actor differentiation.

Since $m^{\mathrm{G}} = R_i/K + K^{-1}\sum_{j\neq i} R_j$, independence gives $\mathbb{E}[R_j Z_i \mid x] = \mathbb{E}[R_j \mid x]\mathbb{E}[Z_i \mid x] = 0$ for each sibling, leaving only the target term and proving Equation 12. For the sibling rule, let $\alpha_x = \mathbb{E}[C_{-i} \mid x]$ and apply conditional independence and integrability to obtain

$$\mathbb{E}[C_{-i}B_{-i}Z_i \mid x] = \mathbb{E}[C_{-i}B_{-i} \mid x]\mathbb{E}[Z_i \mid x] = 0, \tag{13}$$

$$\mathbb{E}[C_{-i}R_i Z_i \mid x] = \mathbb{E}[C_{-i} \mid x]\mathbb{E}[R_i Z_i \mid x] = \alpha_x g_x. \tag{14}$$

Subtracting the two identities proves the theorem, with $\alpha_x > 0$ following from strict positivity of the inverse scale. Although policy parameters affect which sibling statistics are observed, the actor treats the realized values as constants, so the proof evaluates their distribution at fixed parameters without differentiating through the credit calculation. The resulting alignment $\langle \mathbb{E}[\widehat{g}_x \mid x], g_x\rangle = \alpha_x \|g_x\|^2 \geq 0$ concerns the mean over repeated groups and does not fix the sign of any individual update.

### B.3 Comparison with GRPO and RLOO

To compare the credit rules and mean updates in Table 5, define $b_{n,s}(p) = \binom{n}{s}p^s(1-p)^{n-s}$, $m_s = (s + 1/2)/K$, and $c_s = [m_s(1 - m_s)]^{-1/2}$ for the finite sums below.

Table 5: Binary-reward credit rules and expected updates.

| Method | Reference rewards | Baseline | Scale | Credit $A_i$ | Homogeneous credit | | Mean update $\mathbb{E}[\widehat{g}_x \mid x]$ |
|---|---|---|---|---|---|---|---|
| | | | | | $S = 0$ | $S = K$ | |
| GRPO | All $K$ | $m^{\mathrm{G}} = \dfrac{S}{K}$ | $\sigma^{\mathrm{G}} = \sqrt{m^{\mathrm{G}}(1-m^{\mathrm{G}})}$ | $\dfrac{R_i - m^{\mathrm{G}}}{\sigma^{\mathrm{G}}}$ | 0 | 0 | $\gamma_K(p) g_x$ |
| RLOO | Other $K-1$ | $\dfrac{S - R_i}{K-1}$ | 1 | $R_i - \dfrac{S - R_i}{K-1}$ | 0 | 0 | $g_x$ |
| **BLOO** (ours) | Other $K-1$ | $m_{-i} = \dfrac{S - R_i + 1/2}{K}$ | $\sigma_{-i} = \sqrt{m_{-i}(1-m_{-i})}$ | $\dfrac{R_i - m_{-i}}{\sigma_{-i}}$ | $-\dfrac{1}{\sqrt{2K-1}}$ | $+\dfrac{1}{\sqrt{2K-1}}$ | $\alpha_K(p) g_x$ |

Under iid on-policy sampling with detached credits and separate KL, GRPO uses the population-divisor scale and zero credit at $S \in \{0, K\}$; $\alpha_K$ and $\gamma_K$ are defined in Equations 6 and 17.

For a fixed sibling count, a binary target has only two possible credit values, whose difference determines the expected update even when the rule is not expressed as a sibling baseline, thus providing a common comparison of BLOO, GRPO, and RLOO.

**Lemma 1** (Target-reward contrast)**.** *If a credit depends on the target and siblings only through $A_i = f(R_i, S_{-i})$, with finite moments, then*

$$\mathbb{E}[A_i Z_i \mid x] = \sum_{s=0}^{K-1} b_{K-1,s}(p)\,[f(1,s) - f(0,s)]\,g_x. \tag{15}$$

*Proof.* Conditioning on $S_{-i} = s$, which is independent of the target, allows us to write $f(R_i, s) = f(0,s) + [f(1,s) - f(0,s)]R_i$ and apply Equation 11. Averaging over the binomial sibling count gives the stated identity, which exchangeability then transfers to the group average. □

The contrast equals one for RLOO, $(K-1)/K$ for unscaled leave-in centering, and $c_s$ for BLOO, while GRPO assigns the credits below because a successful target gives total count $s+1$ and a failed target gives count $s$.

$$f_{\mathrm{G}}(1,s) = \sqrt{\frac{K-1-s}{s+1}}, \qquad f_{\mathrm{G}}(0,s) = -\sqrt{\frac{s}{K-s}}. \tag{16}$$

$$\mathbb{E}[\widehat{g}^{\mathrm{G}}_x \mid x] = \gamma_K(p) g_x, \quad \gamma_K(p) = \sum_{s=0}^{K-1} b_{K-1,s}(p) \left[\sqrt{\frac{K-1-s}{s+1}} + \sqrt{\frac{s}{K-s}}\right] > 0, \tag{17}$$

These expressions implement zero credit at the homogeneous endpoints and have positive contrast for every $s$, proving Equation 17. Using a sample-divisor standard deviation multiplies GRPO by $\sqrt{(K-1)/K}$, while adding a positive denominator epsilon changes its count coefficients but preserves positive contrast. The resulting binary-reward identities do not extend automatically to arbitrary real rewards, trajectory-dependent scaling, or off-policy updates.

### B.4 Bayesian construction, predictive variance, and fixed-policy moments

To construct the reference, the working model draws $U \sim \mathrm{Beta}(1/2, 1/2)$ followed by conditionally independent binary rewards with success probability $U$, separately from the fixed-policy sampling law with probability $p$. A Bernoulli observation has Fisher information $I(u) = 1/[u(1-u)]$, so Jeffreys' prior density is proportional to $\sqrt{I(u)}$. Multiplying this density by the sibling likelihood $u^{S_{-i}}(1-u)^{K-1-S_{-i}}$ adds the observed counts to the two half-counts, yielding the posterior and prediction below with posterior parameters summing to $K$.

$$\begin{aligned} &U \mid R_{-i} \sim \mathrm{Beta}\big(S_{-i} + \tfrac{1}{2}, K - 1 - S_{-i} + \tfrac{1}{2}\big)\,, \\ &\widetilde{R}_i \mid R_{-i} \sim \mathrm{Bernoulli}(m_{-i}), \qquad m_{-i} = \frac{S_{-i} + 1/2}{K}, \quad \sigma^2_{-i} = m_{-i}(1 - m_{-i}). \end{aligned} \tag{18}$$

To distinguish variation in the next reward from uncertainty about its underlying probability, write $m = m_{-i}$ and apply the law of total variance, which separates the predictive variance into the two components below.

$$\mathrm{Var}(\widetilde{R}_i \mid R_{-i}) = \mathbb{E}[U(1-U) \mid R_{-i}] + \mathrm{Var}(U \mid R_{-i}) = \frac{K}{K+1} m(1-m) + \frac{1}{K+1} m(1-m). \tag{19}$$

Thus the predictive variance is $m(1-m)$ and the posterior parameter variance is $m(1-m)/(K+1)$, giving $(\widetilde{R}_i - m)/\sigma_{-i}$ conditional mean zero and variance one under the hierarchical model. Moreover, $\mathbb{E}[(\widetilde{R}_i - a)^2 \mid R_{-i}] = m(1-m) + (a-m)^2$ makes $m$ the unique squared-error-optimal reward prediction, although this optimality does not imply that it minimizes the covariance of $A_i Z_i$.

For fixed policy success probability $p$, the held-out reward instead remains $\mathrm{Bernoulli}(p)$ independently of the siblings, yielding the fixed-policy moments below.

$$\mathbb{E}[A^{\mathrm{B}}_i \mid S_{-i} = s, x] = \frac{p - m_s}{\sqrt{m_s(1-m_s)}}, \quad \mathbb{E}[(A^{\mathrm{B}}_i)^2 \mid x] = \sum_{s=0}^{K-1} b_{K-1,s}(p) \frac{p(1-p) + (p-m_s)^2}{m_s(1-m_s)}. \tag{20}$$

A predictively centered rule therefore need not have zero conditional mean for every fixed policy and observed sibling count, nor does predictive centering require re-centering the realized credits within a group.

**Other symmetric priors.** A $\mathrm{Beta}(a, a)$ prior with fixed $a > 0$ gives $m_{-i}^{(a)} = (S_{-i} + a)/(K - 1 + 2a)$, with prediction and scale determined by siblings and the target reward entering only through the standardized error. The resulting credit is symmetric and bounded by $\sqrt{(K - 1 + a)/a}$, with homogeneous values $\pm\sqrt{a/(K - 1 + a)}$, making Jeffreys' $a = 1/2$ one admissible smoothing rule within this family. We choose it for the invariance of its prior measure under smooth one-to-one reparameterizations, which does not require its density to be constant in every coordinate system. This invariance motivates the default without asserting minimum gradient variance, and predictive standardization describes the working reward model rather than the trained policy's reported confidence.

### B.5 Proof of finite bounds and closed-form credit

To prove Proposition 1, we bound the prediction away from zero and one and substitute each possible target reward, obtaining a pointwise credit bound that also controls the weighted score's second moment in Equation 21.

$$\mathbb{E}[\|A_i^{\mathrm{B}} Z_i\|_2^2 \mid x] \le (2K - 1)\mathbb{E}[\|Z_i\|_2^2 \mid x]. \tag{21}$$

Since $0 \le S_{-i} \le K - 1$, the posterior mean lies in $[1/(2K), 1 - 1/(2K)]$, giving the following pointwise bounds on its inverse scale.

$$2 \le [m(1 - m)]^{-1/2} \le \frac{2K}{\sqrt{2K - 1}}. \tag{22}$$

The successful and failed target credits, $A_i^{\mathrm{B}} = \sqrt{(1 - m)/m}$ and $-\sqrt{m/(1 - m)}$, respectively, yield Proposition 1 at the endpoints and Equation 21 after squaring, while substitution of the prediction at total count $s$ gives the closed forms below.

$$a_1(s) = \sqrt{\frac{K - s + 1/2}{s - 1/2}}\ (1 \le s \le K), \qquad a_0(s) = -\sqrt{\frac{s + 1/2}{K - s - 1/2}}\ (0 \le s \le K - 1). \tag{23}$$

These expressions need only be defined for outcomes present in the group and preserve reward symmetry, as permuting trajectories permutes their credits and complementing all rewards negates them. A unique success attains $\sqrt{2K - 1}$, while each failure in the same group receives $-\sqrt{3/(2K - 3)}$.

Table 6: Exact BLOO credits at $K = 5$, rounded to three decimals. The group mean is intentionally not constrained to zero.

| $S$ | Success credit | Failure credit | Group mean | GRPO active? |
|---|---|---|---|---|
| 0 | – | −0.333 | −0.333 | No |
| 1 | 3.000 | −0.655 | 0.076 | Yes |
| 2 | 1.528 | −1.000 | 0.011 | Yes |
| 3 | 1.000 | −1.528 | −0.011 | Yes |
| 4 | 0.655 | −3.000 | −0.076 | Yes |
| 5 | 0.333 | – | 0.333 | No |

### B.6 The exact multiplier and transformed objective

The mean identity reduces the calculation to the expected inverse scale, which we obtain by averaging $c_s = [m_s(1 - m_s)]^{-1/2}$ over $S_{-i} \sim \mathrm{Binomial}(K - 1, p)$, giving Equation 6. The multiplier is a continuous polynomial whose bounds follow from those on $c_s$, while the relation $c_{K-1-s} = c_s$ establishes the symmetry $\alpha_K(p) = \alpha_K(1 - p)$.

Writing $I_p(a, b)$ for the regularized incomplete beta function allows us to express the potential in the following finite closed form.

$$\Psi_K(p) = \frac{1}{K}\sum_{s=0}^{K-1} c_s I_p(s + 1, K - s), \qquad \Psi_K(1) = \frac{1}{K}\sum_{s=0}^{K-1} c_s. \tag{24}$$

Integrating each term $\binom{K-1}{s}u^s(1-u)^{K-1-s}$ gives $I_p(s+1, K-s)/K$, so the potential's derivative is $\alpha_K(p) \geq 2$, and the chain rule together with the prompt-expectation interchange proves Theorem 2. Although the actor treats credits as fixed, their expected score field already equals this gradient, so differentiating through the credit computation would change the analyzed estimator.

**Why per-prompt alignment is insufficient.** To show how prompts can favor opposite changes in shared parameters, consider equally likely prompts with $p_1(\theta) = 0.5 + \theta$ and $p_2(\theta) = 0.99 - 0.8\theta$, both valid smooth Bernoulli policies for scalar $\theta$ near zero. Raw mean reward has derivative $(1 - 0.8)/2 = 0.1 > 0$, so increasing $\theta$ improves that objective, whereas BLOO at $K = 5$ has $\alpha_5(0.5) \simeq 2.258$ and $\alpha_5(0.99) \simeq 3.288$, giving $F_5'(0) \simeq -0.186 < 0$. The stronger weight on the second prompt therefore reverses the aggregate direction, even though each prompt's expected update remains aligned with its own reward gradient.

### B.7 Homogeneous signal and its relative size

To prove Proposition 2, we calculate the all-success and all-failure contributions separately, with the target's reward-weighted score already included in $g_x$ and only the $K-1$ siblings supplying a probability factor. The mean coefficient therefore involves powers of $K-1$, whereas the frequency $q_K$ of the full homogeneous event involves powers of $K$.

On all-success groups, credit is $1/\sqrt{2K-1}$ and the siblings succeed independently with probability $p^{K-1}$, yielding a contribution of $p^{K-1}g_x/\sqrt{2K-1}$ to $\mathbb{E}[A_iZ_i \mid x]$. On all-failure groups, combining $\mathbb{E}[(1-R_i)Z_i \mid x] = -g_x$ with negative credit gives $(1-p)^{K-1}g_x/\sqrt{2K-1}$, so adding the two contributions and averaging over $i$ proves the first two parts of Equation 8. To bound their share, retain the two endpoint terms in the binomial sum for $\alpha_K(p)$, giving

$$\alpha_K(p) \geq c_0\{p^{K-1} + (1-p)^{K-1}\} = 2K\, h_K(p), \tag{25}$$

because all remaining terms are nonnegative, proving the $1/(2K)$ fraction bound. When $K = 2$, no interior terms remain and equality holds everywhere; when $K > 2$ and $0 < p < 1$, the inequality is strict and approaches equality at the endpoints. At an interior parameter point with $p = 0$ or 1, differentiability and the probability constraint imply $g_x = 0$, so nonzero credits need not yield a nonzero mean update.

### B.8 Proof of the group variance decomposition and a boundary regime

Using the credits, covariance $\Sigma_r$, and multiplier $u$ defined before Theorem 3, we fix the reward vector and then average over its success count to separate the noise sources while preserving credit dependence, yielding the homogeneous within-outcome contribution in Equation 26.

$$V_{H,\text{within}} = \frac{p^K \operatorname{tr}\Sigma_1 + (1-p)^K \operatorname{tr}\Sigma_0}{K(2K-1)}. \tag{26}$$

For $0 < p < 1$, the score identities imply

$$\mu_1 := \mathbb{E}[Z_i \mid R_i = 1, x] = g_x/p, \qquad \mu_0 := \mathbb{E}[Z_i \mid R_i = 0, x] = -g_x/(1-p). \tag{27}$$

Conditioning on the entire reward vector preserves trajectory independence under the respective reward-conditional laws, and fixing $S = s$ determines the credits and outcome counts, giving

$$\mathbb{E}[\widehat{g}_x \mid S = s, x] = u(s;p)g_x, \tag{28}$$

$$\operatorname{Cov}(\widehat{g}_x \mid S = s, x) = K^{-2}\{s\, a_1(s)^2\Sigma_1 + (K-s)a_0(s)^2\Sigma_0\}. \tag{29}$$

Conditioning only on $S$ yields the same covariance because every ordering with that count has the same conditional mean and covariance. Applying the law of total covariance over $S \sim \text{Binomial}(K, p)$ and taking the trace then proves Theorem 3, while preserving the shared-group dependence that would be lost by treating the weighted scores $A_iZ_i$ as independent. Substituting BLOO's homogeneous credits into the second component gives Equation 26, quantifying the noise retained within those groups.

**Why no uniform variance improvement is possible.** The four-action bandit in Section 5.6 has $\operatorname{tr}\Sigma_0 = 1 - 0.6^2 - 0.4^2 = 0.48$ and $\operatorname{tr}\Sigma_1 = 1 - 0.7^2 - 0.3^2 = 0.42$, so as $p \downarrow 0$, the homogeneous part of BLOO's *scale-normalized* within-outcome variance tends to

$$\frac{0.48}{K(2K-1)}\left(\frac{\sqrt{2K-1}}{2K}\right)^2 = \frac{0.48}{4K^3} > 0. \tag{30}$$

Since $\|g_x\|^2 = 1.10\,p^2(1-p)^2$, BLOO's normalized relative MSE diverges at least as $p^{-2}$ in this regime. GRPO and RLOO instead vanish on homogeneous groups and have nonzero-group probability $O(p)$, bounded fixed-$K$ scores and credits, and multipliers with finite positive limits, giving relative MSE at most $O(p^{-1})$. BLOO is therefore noisier sufficiently near the boundary despite its mean identities, while the specific noise source disappears when each reward class contains only one action and $\Sigma_0 = \Sigma_1 = 0$.

### B.9 Population limit and its scope

Because the prior has fixed total weight, its influence diminishes as more siblings are observed, as the following result establishes by holding the target reward and an interior success probability fixed while increasing the number of independent siblings.

**Proposition 3** (Coupled interior limit)**.** *Fix $p \in (0,1)$ and an infinite iid sequence $R_0, R_1, \ldots \sim \mathrm{Bernoulli}(p)$. Hold out $R_0$ and set $m_K = (1/2 + \sum_{j=1}^{K-1} R_j)/K$. Then*

$$m_K \to p \text{ a.s.}, \qquad \frac{R_0 - m_K}{\sqrt{m_K(1-m_K)}} \to \frac{R_0 - p}{\sqrt{p(1-p)}} \text{ a.s.}, \qquad \alpha_K(p) \to [p(1-p)]^{-1/2}. \tag{31}$$

*Proof.* The strong law and vanishing prior weight imply $m_K \to p$, after which continuity in an interior neighborhood establishes convergence of the held-out credit. To pass to the expected inverse scale, choose $\delta > 0$ with $[p - 2\delta, p + 2\delta] \subset (0,1)$, so the inverse scale is eventually uniformly bounded on $|\widehat{p}_{K-1} - p| \le \delta$ and dominated convergence applies there. On the complement, the inverse scale is at most $2K/\sqrt{2K-1}$ and Hoeffding's bound gives probability at most $2e^{-2(K-1)\delta^2}$, whose product vanishes, proving convergence of the full expectation. The argument requires fixed interior $p$ and does not establish a uniform limit when $p$ approaches a boundary with $K$. □

## C Implementation and Verification Details

The source includes `bloo_estimator.py`, a minimal NumPy credit implementation that accepts shape $(\ldots, K)$, checks $K \ge 2$ and finite binary rewards in $\{0,1\}$, and returns floating-point credits. To integrate it into a trainer, group responses by prompt, compute counts and scales in FP32 or higher, and detach the credits before applying them to valid response-token sums that include sampled termination tokens and exclude padding. Keep the credits floating point and uncentered, with token-level KL separate from the binary verifier reward, to preserve the estimator's definition.

To connect the estimator with a practical PPO-style actor objective (Schulman et al., 2017), draw $x \sim \mathcal{D}$ and $K$ iid responses from $\pi_{\theta_{\text{old}}}(\cdot \mid x)$, denoting this expectation by $\mathbb{E}_{\text{old}}$ and detaching the computed response credits. For valid response length $T_i$ and token ratio $\rho_{i,t}(\theta) = \pi_\theta(Y_{i,t} \mid x, Y_{i,<t}) / \pi_{\theta_{\text{old}}}(Y_{i,t} \mid x, Y_{i,<t})$, maximize

$$\mathcal{L}_{\mathrm{B}}(\theta) = \mathbb{E}_{\text{old}}\left[\frac{1}{K}\sum_{i=1}^{K}\sum_{t=1}^{T_i} \min\left\{\rho_{i,t}(\theta)\,A_i^{\mathrm{B}}, \operatorname{clip}(\rho_{i,t}(\theta), 1-\epsilon_\ell, 1+\epsilon_h)\,A_i^{\mathrm{B}}\right\}\right] - \beta\mathcal{R}_{\mathrm{KL}}(\theta). \tag{32}$$

For the surrogate in Equation 32, one precise reference penalty is

$$\mathcal{R}_{\mathrm{KL}}(\theta) = \mathbb{E}_{x,Y\sim\pi_{\theta_{\text{old}}}} \sum_{t=1}^{T} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x, Y_{<t}) \,\|\, \pi_{\text{ref}}(\cdot \mid x, Y_{<t}))\,. \tag{33}$$

At the rollout parameter, token ratios and their clipped values equal one, with clipping locally inactive for positive clip widths, so the token-sum policy term differentiates to the sequence-score estimator

studied in the theory and KL adds a separate gradient. Dividing the policy term by a fixed common length constant only rescales it, whereas dividing by each sampled length or the random batch token total can change its mean direction. Alternative KL estimators, policy reuse, and reward-based filtering therefore require explicit specification and further analysis of their combined effects.

**Exact task specification.** For a categorical softmax policy with probabilities $\boldsymbol{q}$, the action-$j$ score is $e_j - \boldsymbol{q}$ and its reward-gradient component is $g_{x,j} = q_j(r_j - p)$. The two-action task uses $\boldsymbol{q} = (p, 1-p)$ with rewards $(1, 0)$, while the four-action task uses $\boldsymbol{q} = (0.7p, 0.3p, 0.6(1-p), 0.4(1-p))$ with rewards $(1, 1, 0, 0)$ to introduce within-outcome diversity at the same success probability. To match mean strength, set $\widehat{g}_x^{\mathrm{norm}} = \widehat{g}_x / w_K(p)$, where $w_K$ is $\gamma_K$, 1, or $\alpha_K$ for GRPO, RLOO, or BLOO, and compute relative MSE as $\mathbb{E}\|\widehat{g}_x^{\mathrm{norm}} - g_x\|_2^2 / \|g_x\|_2^2$.

**Exact numerical verification.** `verify_theory.py` checks the binary credit identities over exhaustive ordered reward groups, then independently constructs categorical score updates for each unordered action-count vector with probability $K! \prod_j q_j^{n_j} / \prod_j n_j!$, preserving the weights of all ordered groups. The script verifies the mean multiplier, homogeneous contribution, and group-variance decomposition, checks the potential derivative by finite differences, and saves unrounded results together with both variance components. These finite-task calculations are exact up to floating-point rounding and provide mathematical verification without constituting language-model training measurements.

# D Additional Evaluation Details

This appendix gives the protocol and supplementary comparisons for Section 5.

## D.1 Evaluation Protocol and Data Sources

**Matched training.** Visual comparisons follow Section 5.1 and share the SFT initialization, prompt stream, decoding settings, generated-token budget, optimizer, KL regularization, and loss reduction, with $K = 5$. The supplementary text scenarios use Qwen3-1.7B and Qwen3-8B on DAPO-Math-17K (Yu et al., 2025), with $K = 8$. The protocol uses five training seeds for the visual and 1.7B models and three for 8B.

**Observed baseline and scores.** The supplied Qwen3-VL-4B baseline covers Geometry3K, MathVista, MathVision, MathVerse, We-Math, MMMU, and MMStar (Lu et al., 2021; 2023; Wang et al., 2024a; Zhang et al., 2024; Qiao et al., 2025; Yue et al., 2024; Chen et al., 2024). Its split labels are test for Geometry3K/MathVision, testmini for MathVista/MathVerse/We-Math, validation for MMMU, and val for MMStar. Accuracy is correct/completed, with an unweighted seven-task mean of 67.1%. All averages and gains use unrounded values.

**Evaluation and uncertainty.** The matched evaluation uses Mean@8, averaging correctness over eight independent responses per prompt. Four evaluation seeds give 32 responses per prompt and checkpoint. The original model requires reevaluation because its supplied aggregates do not specify the sampling protocol, checkpoint revision, or MathVerse variant composition. Training-seed variation and evaluation-sampling uncertainty are reported separately. Paired confidence intervals use a hierarchical bootstrap over training seeds and prompts, grouping related MathVerse variants and We-Math subproblems by source problem. Final-checkpoint and best-checkpoint results are distinguished.

**Component and optimization controls.** Predictive scaling is tested with learning-rate sweeps and matched update norms. RLOO, BNPO, and EBPO provide centering, normalization, and shrinkage controls (Ahmadian et al., 2024; Xiao et al., 2025; Han et al., 2026); EBPO's global-statistics update is evaluated separately. Full-group Jeffreys tests target inclusion, while homogeneous masking preserves all mixed-group credits. Additional controls compare Jeffreys and uniform priors, token-mean and sequence-mean losses, and clipping, KL, and rollout reuse. SIGNBALANCE and breakdowns by answer-space size test sensitivity to guessing on bounded-answer tasks (Wang et al., 2026b).

**Budgets and trajectory metrics.** Generated tokens, actor-training tokens, and wall time are recorded separately because filtering and adaptive allocation can discard generated responses (Yu et al., 2025; Zhang et al., 2026). Scalar-credit overheads are not end-to-end measurements. Component trajectories share the SFT start of 62.4%; AUC integrates accuracy over normalized progress, the last-20% mean covers its final fifth, and drop is peak minus final accuracy. Twice-horizon scores evaluate training with twice the standard budget. Gradient norms are recorded before and after clipping, with policy and KL terms separated; entropy and response length provide further diagnostics. Filtering, adaptive stopping, and rollout reuse fall outside the fixed iid sampling assumptions of the exact analysis.

### D.2 Estimator and System Comparisons

Table 7 separates changes to scalar credit from changes to sampling or sequence-level optimization. In the estimator comparison, BLOO exceeds DBB by 0.8 and 0.7 pp on the 1.7B and 8B models. Replacing GRPO credit within GSPO gives gains of 1.5 and 1.4 pp, testing whether BLOO can complement a different policy optimizer. The Dr.GRPO row uses unscaled full-group centering under the common loss; its full normalization scheme requires a separate system comparison. Relative cost is included to assess score improvements alongside resource use.

Table 7: Text-model comparisons. Math avg. covers six in-domain benchmarks; cost is relative to GRPO GPU-hours.

| Estimator | 1.7B math avg. | 8B math avg. | Relative cost |
|---|---|---|---|
| GRPO | 26.6 | 49.0 | 1.000 |
| RLOO | 27.4 | 49.5 | 1.000 |
| Dr.GRPO | 28.0 | 49.7 | 0.998 |
| DBB | 30.0 | 51.3 | 1.002 |
| BLOO | **30.8** | **52.0** | 1.002 |
| DAPO (budget matched) | 29.2 | 50.6 | 1.18 |
| GRESO (budget matched) | 29.4 | 50.8 | 0.84 |
| GSPO + GRPO credit | 29.8 | 51.1 | 1.01 |
| GSPO + BLOO credit | **31.3** | **52.5** | 1.01 |

### D.3 Homogeneous-Group Credit

Table 8 isolates the feedback lost when all rewards agree. BLOO gives nonzero credit to every homogeneous token, whereas GRPO and the masked variant give none. The comparison tests whether using these already-generated responses improves learning without additional rollouts. Because weighted response gradients can cancel, nonzero credit alone does not establish a nonzero or useful group gradient.

Table 8: Homogeneous-group credit at $K = 5$. Token use is the analytic fraction of homogeneous tokens with nonzero credit.

| Method | Token use | $S = 0$ grad norm | $S = K$ grad norm | Accuracy (%) |
|---|---|---|---|---|
| GRPO | 0% | 0 | 0 | 65.5 |
| BLOO, groups masked | 0% | 0 | 0 | 67.4 |
| Full BLOO | **100%** | .31 | .24 | **68.7** |

### D.4 History and Curriculum Shifts

Table 9 tests DBB with fixed and validation-selected discounts (Kim et al., 2026), using its documented history initialization and recording discount-search cost. In this scenario, tuning DBB reduces recovery time from 34 to 28 updates, while BLOO requires 11 updates and a higher post-shift AUC (.716 versus .681). The smaller repeated-to-unseen drop for BLOO (0.5 versus 1.6 pp) further tests the value of current-group evidence when prompts or difficulty change. Recovery time and AUC capture the speed and sustained quality of adaptation, respectively.

Table 9: Qwen3-1.7B adaptation. Recovery counts updates needed to regain the pre-shift normalized reward.

| Method | Repeated | No-repeat | Post-shift AUC | Recovery ↓ |
|---|---|---|---|---|
| DBB, global $\lambda = .5$ | 29.5 | 27.9 | .663 | 34 |
| DBB, searched $\lambda$ | 30.0 | 28.4 | .681 | 28 |
| BLOO, no tuning | **30.8** | **30.3** | **.716** | **11** |

### D.5 Model Scale and Transfer

Figure 3(c) compares in-domain (ID) and out-of-domain (OOD) gains. Text ID averages MATH500, Minerva, AIME24, AIME25, AMC24, and OlympiadBench (Hendrycks et al., 2021; He et al., 2024); text OOD averages MMLU-Pro, GPQA-Diamond, and BBH (Wang et al., 2024b; Rein et al., 2023; Suzgun et al., 2023). Visual ID uses Geometry3K, and visual OOD averages MathVista and MathVision (Lu et al., 2023; Wang et al., 2024a). The other four visual benchmarks remain separately reported in Table 1. This breakdown tests whether gains extend beyond the training task and across model scales.

## E Limitations and Broader Impact

The exact theory assumes a fixed binary verifier and conditionally independent on-policy responses, so fractional rewards, correlated decoding, reward-based filtering, length normalization, clipping after policy movement, KL regularization, and policy reuse can alter the analyzed update. The prior supplies a working prediction model without asserting that it describes the true distribution of prompt difficulty.

Because the credit bound grows as $\sqrt{K}$ and does not control score norms, it cannot guarantee lower gradient variance, while retained homogeneous credit and prompt reweighting can respectively introduce noise and conflicts under shared parameters. Verifier-sensitive checks also remain necessary because reinforcing false positives can increase reward without establishing more reliable reasoning.

The finite tasks verify estimator identities and the supplied ViRL39K traces provide initial accuracy observations, but neither establishes multi-seed evidence of general benchmark gains, transfer, robustness, or throughput. The main comparisons identify these outstanding experiments and require results before they can support broader empirical claims.

## F Related Work

**Group-based optimization and its objective.** Group-based methods alter how rewards enter optimization through length and reward normalization in Dr.GRPO, sampling and loss changes in DAPO, and sequence-level importance ratios and clipping in GSPO (Liu et al., 2025; Yu et al., 2025; Zheng et al., 2025a). Local-curvature analysis identifies conditions under which standard-deviation normalization accelerates optimization (Ge et al., 2026), while objective-focused work develops leave-two-out estimators for max@$K$ in MaxPO (Takashiro et al., 2026) and shows how pass@$k$ prompt weighting can conflict with pass@1 under shared parameters (Barakat et al., 2026). BLOO's analysis identifies its transformation of mean verifier reward, thus making explicit the distinction between a correct per-prompt direction and the aggregate objective followed during training.

**Bayesian references and finite-group information.** Prior work uses several sources of statistical evidence, with BNPO fitting a Beta-density normalizer (Xiao et al., 2025), EBPO shrinking local means toward global statistics (Han et al., 2026), and discounted Beta–Bernoulli estimation accumulating decayed prompt-specific counts (Kim et al., 2026). AERO combines Beta posterior means with rollout allocation and selection (Zhang et al., 2026), GRESO uses past reward dynamics to select prompts before rollout (Zheng et al., 2025b), and AVSPO introduces virtual rewards when group contrast collapses (He et al., 2026). Building on these precedents for smoothing and nonzero homogeneous credit, BLOO constructs both the prediction and next-reward scale from each target's

current siblings and analyzes the resulting mean objective, retained contribution, and full gradient variance.

**Uncertainty and the level of credit assignment.** Uncertainty enters learning at different levels, with GUPO weighting query-level gradients according to their uncertainty (Guo et al., 2026) and OPPO using oracle-conditioned likelihood evidence to form Bayesian token-level credits (Li et al., 2026). BLOO instead derives one scalar credit per complete response from binary sibling outcomes, so adding token-level credit or group weighting would change the estimator and require a corresponding analysis.

**What a binary verifier can establish.** Because binary correctness leaves aspects of reasoning unobserved, SIGNBALANCE addresses reinforced guessing through sign-based magnitudes and class balancing (Wang et al., 2026b), while Prism-GRPO introduces trajectory-quality feedback in vision-language-action tasks (Deng et al., 2026). BLOO retains signal relative to the given verifier, with its variance analysis showing how identical labels can conceal different update directions without inferring reasoning quality from those labels. Reported confidence raises a separate concern, as Bereket & Leskovec (2025) show that GRPO normalization can produce overconfidence for stochastic outcomes, whereas BLOO's predictive moments describe its working reward model rather than the confidence calibration of the trained policy.

## G AI USE STATEMENT

We disclose the use of GPT-6-Astra to assist with language refinement and manuscript preparation, including literature and code analysis, mathematical formulation and proof drafting, interpretation of supplied results, and visualization-script preparation. The measured training traces and evaluation results were supplied by the authors, and AI-assisted analysis did not replace model training or evaluation. The authors retain responsibility for the BLOO method and implementation, verification of mathematical arguments and empirical claims, and all research decisions, as well as the accuracy, originality, and integrity of the final manuscript and accompanying artifacts.