Metric-Projected GRPO: Selecting a Positive-Credit Gauge for Multi-Sample RLVR

Anonymous ACL submission

Abstract

Group Relative Policy Optimization (GRPO), widely adopted in RLVR for reasoning models, has advanced mathematical reasoning without requiring a learned value model. Nevertheless, in mixed binary groups, GRPO leaves a key credit-allocation choice unresolved: reward normalization fixes the total correct-versus-incorrect contrast, but not how positive credit is distributed among observed correct modes. We propose METRIC-PROJECTED GRPO (MP-GRPO), enhancing GRPO with target-K-aware positive-credit redistribution. MP-GRPO aggregates correct rollouts into modes, estimates their local old-policy mass, and applies entropic projection to assign more credit to modes with lower target-K coverage. Experiments under an aligned Qwen3-Base protocol show that MP-GRPO improves averaged Acc@8/Pass@8 over vanilla GRPO by $+4.59/+7.14$ points on Qwen3-1.7B and $+7.45/+6.66$ points on Qwen3-8B across six mathematical reasoning benchmarks.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a widely used post-training paradigm for improving reasoning language models (Lambert et al., 2024; Guo et al., 2025; Hou et al., 2025; Xie et al., 2025). In mathematical reasoning, code generation, and formal proof search, models can sample candidate solutions and learn from automatically checkable outcomes. Group Relative Policy Optimization (GRPO) is particularly attractive in this setting because it avoids training a separate value model and estimates advantages from rollout groups sampled for the same prompt (Shao et al., 2024). As multi-sample inference becomes common, however, reasoning performance depends not only on improving the correctness of one sampled answer, but also on covering distinct correct solution modes within a finite sampling budget.

Based on where they intervene in the RLVR pipeline, existing methods can be roughly categorized into three classes. (1) Optimizer-level methods refine group-relative training through sampling rules, clipping strategies, normalization, or sequence-level ratio design (Yu et al., 2025; Zheng et al., 2025; Yang et al., 2026). (2) Distillation and entropy-aware methods improve on-policy transfer or preserve plausible high-entropy reasoning behavior during post-training (Lu and Lab, 2025; Jin et al., 2026). (3) Multi-sample and correct-support methods modify the outer reward, batch-level transformation, or correct-solution distribution to better align training with Pass@K-style evaluation (Chen et al., 2025; Walder and Karkhanis, 2026; Peng et al., 2025; Plyusov et al., 2026; Lochab et al., 2026). These approaches demonstrate that optimization design and correct-solution support are both important for multi-sample reasoning. Nevertheless, they have not isolated the allocation choice that remains after a mixed binary-GRPO group has already been sampled and verified: how should the fixed positive update mass be distributed among the observed correct rollouts?

To fill this gap, we study the within-correct credit-allocation problem in binary group-relative RLVR. Although one may directly apply existing Pass@K-oriented or diversity-oriented objectives, two challenges remain. **First, fixed contrast does not imply fixed allocation.** In a sampled mixed binary group, the group statistics fix the total contrast between correct and incorrect rollouts but do not specify the distribution of positive credit within the correct set. Standard GRPO implicitly uses a uniform allocation over observed correct rollouts, but this is only one possible contrast-preserving gauge. **Second, exact Pass@K does not determine this gauge.** It depends on the total probability mass assigned to the full correct set, not on how that mass is distributed among distinct correct modes. Thus, Pass@K encourages increasing total correct

probability, but it does not by itself select a non-uniform direction for routing positive credit among observed correct modes.

To address the above challenges, we present METRIC-PROJECTED GRPO (MP-GRPO), a minimal intervention for selecting a target-K-aware positive-credit gauge. The method preserves the verifier reward, rollout budget, clipped GRPO surrogate, and group-level contrast, and changes only where the already-assigned positive credit flows. Specifically, MP-GRPO consists of three steps. First, because duplicated rollouts can express the same solution, it aggregates observed correct rollouts into modes. Second, it estimates each mode's local correct-conditional old-policy mass and computes a target-K support-coverage marginal, which identifies modes that are unlikely to be covered under the intended sampling budget. Third, it performs an entropic projection from the uniform GRPO gauge to obtain a smooth coverage-aware gauge, thereby redirecting positive credit toward under-covered modes without changing the total positive mass.

To sum up, the contributions of this work are as follows:

- We formulate the within-correct credit-allocation freedom in binary group-relative RLVR as a positive-credit gauge under fixed group-level contrast, and show that exact Pass@K does not determine this gauge.
- We propose METRIC-PROJECTED GRPO (MP-GRPO), a contrast-preserving method that selects a target-K-aware gauge by combining observed-mode aggregation, support-coverage marginals, and entropic projection.
- We conduct experiments under an aligned Qwen3-Base reasoning protocol. MP-GRPO improves averaged Acc@8/Pass@8 over vanilla GRPO by $+4.59/+7.14$ points on Qwen3-1.7B and by $+7.45/+6.66$ points on Qwen3-8B across six mathematical reasoning benchmarks.

Anonymous code. An anonymous implementation is available at `https://anonymous.4open.science/r/MP-GRPO-CF8F`.

2 Related Work

Group-relative RLVR and multi-sample objectives. Verifiable rewards are now widely used for reasoning post-training in mathematical and rule-checkable domains (Shao et al., 2024; Guo et al., 2025; Lambert et al., 2024; Xie et al., 2025). GRPO is a representative group-relative method because it avoids a learned value model and normalizes rewards within prompt-level rollout groups (Shao et al., 2024). Recent work improves this pipeline through optimizer-level changes such as sampling, clipping, normalization, and sequence-level ratios (Yu et al., 2025; Zheng et al., 2025; Yang et al., 2026), or through objectives and distillation procedures better aligned with multi-sample reasoning (Lu and Lab, 2025; Jin et al., 2026; Chen et al., 2025; Walder and Karkhanis, 2026; Peng et al., 2025). These methods change the optimizer, outer objective, or training signal. In contrast, METRIC-PROJECTED GRPO keeps the verifier, rollout budget, clipped GRPO surrogate, and group-level contrast fixed, and only selects the positive-credit gauge inside a sampled mixed binary group.

Correct-solution support and credit redistribution. Closer to our motivation are works showing that reasoning RL can lose useful support by over-concentrating on dominant trajectories. Entropy- and minority-behavior analyses connect effective RL to high-entropy or underrepresented reasoning signals (Wang et al., 2025; Cheng et al., 2025; Cui et al., 2025; Agarwal et al., 2025). Rarity, Pass@K, and conditional-uniformity methods explicitly encourage rare or balanced correct solutions (Plyusov et al., 2026; Lochab et al., 2026; Peng et al., 2025). Budget-allocation and curriculum methods instead decide where to spend rollout computation across prompts or samples (Shi et al., 2025; Zeng et al., 2025; Li et al., 2025; Yao et al., 2026). These studies establish that support coverage matters, but they either impose rarity/uniformity priors, alter rollout allocation, or optimize an outer multi-sample objective. METRIC-PROJECTED GRPO instead views within-correct allocation as an underdetermined contrast-preserving gauge and selects a target-K-aware credit direction over observed correct modes.

3 Problem Scope

In this section, we formalize the problem scope of MP-GRPO: the positive-credit allocation left open by binary GRPO after sampling and verification. We first rewrite binary GRPO as a fixed correct-vs.-incorrect contrast plus a selectable within-correct gauge. We then show that exact Pass@K does not

determine this gauge, motivating a separate credit-allocation rule for finite-budget multi-sample reasoning.

3.1 Binary GRPO as Contrast Plus Gauge

For a prompt x, the old policy $\pi_{\theta_{\text{old}}}$ samples a group of G rollouts $\{\tau_i\}_{i=1}^{G}$, where $\tau_i = (y_{i,1}, \ldots, y_{i,L_i})$ has length L_i. A verifier returns a binary reward $r_i \in \{0, 1\}$. Standard GRPO normalizes rewards within the group:

$$\bar{r}_x = \frac{1}{G}\sum_{i=1}^{G} r_i, \qquad \sigma_x = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \bar{r}_x)^2}, \tag{1}$$

and uses the normalized advantage

$$A_i^{\text{GRPO}} = \frac{r_i - \bar{r}_x}{\sigma_x + \epsilon_{\text{std}}}, \tag{2}$$

where ϵ_{std} is used for reward normalization and is distinct from the clipping range ϵ_{clip} used in the base clipped GRPO objective. This sequence-level advantage is broadcast to all tokens of rollout i in the clipped surrogate.

Let $C_x = \{i : r_i = 1\}$ and $W_x = \{i : r_i = 0\}$ denote the observed correct and incorrect rollout indices in this sampled group, with $n_x^+ = |C_x|$ and $n_x^- = |W_x|$. These are observed index sets, not full output-space solution sets. For a mixed group, $0 < n_x^+ < G$, binary GRPO assigns only two advantage values:

$$A_i^{\text{GRPO}} = \begin{cases} \Gamma_x / n_x^+, & i \in C_x, \\ -\Gamma_x / n_x^-, & i \in W_x, \end{cases} \tag{3}$$

where

$$\Gamma_x = \frac{n_x^+ n_x^-}{G(\sigma_x + \epsilon_{\text{std}})} \tag{4}$$

is the group-level contrast magnitude. Thus, the observed correct set receives total positive mass Γ_x, while the observed incorrect set receives total negative mass $-\Gamma_x$.

The fixed contrast does not force a uniform split inside the correct set. Let $q^+ \in \Delta(C_x)$ and $q^- \in \Delta(W_x)$ be distributions over the observed correct and incorrect indices, where $\Delta(S)$ denotes the probability simplex over a finite set S. The contrast-preserving family

$$A_i(q^+, q^-) = \begin{cases} \Gamma_x q_i^+, & i \in C_x, \\ -\Gamma_x q_i^-, & i \in W_x \end{cases} \tag{5}$$

preserves the same total positive and negative masses for any normalized q^+ and q^-. Standard GRPO corresponds to one particular gauge: $q_i^+ = 1/n_x^+$ and $q_i^- = 1/n_x^-$. Therefore, the uniform split is not forced by the binary reward. It is a contrast-preserving allocation choice.

MP-GRPO uses the positive side of this nullspace and keeps the negative side uniform. The target coverage signal is defined over verified correct modes, whereas incorrect rollouts do not provide mode-level correct support. For groups with $n_x^+ = 0$ or $n_x^- = 0$, MP-GRPO follows the base GRPO handling and applies no projected-gauge correction.

3.2 Pass@K Does Not Determine the Gauge

A natural question is whether Pass@K itself resolves this allocation freedom. It does not. We use C_x for the observed correct indices in one sampled group and $\mathcal{C}_x$ for the full set of correct output sequences for prompt x. Define the total correct probability

$$P_C = \sum_{\tau \in \mathcal{C}_x} \pi_\theta(\tau \mid x). \tag{6}$$

Exact Pass@K is

$$P_{\text{pass@}K}(x) = 1 - (1 - P_C)^K. \tag{7}$$

It depends only on the total probability of sampling any correct output, not on how that probability is distributed among correct outputs. For any correct output $\tau \in \mathcal{C}_x$ with $p_\tau = \pi_\theta(\tau \mid x)$,

$$\frac{\partial P_{\text{pass@}K}}{\partial p_\tau} = K(1 - P_C)^{K-1}. \tag{8}$$

The marginal value is identical for all correct outputs. Exact Pass@K therefore encourages increasing total correct probability, but gives no non-uniform direction for allocating positive credit within the correct set. The MP-GRPO gauge is not the derivative of exact Pass@K, but a target-aware tie-breaker among contrast-preserving gauges that exact Pass@K leaves underdetermined.

4 Methodology

To address the allocation freedom identified above, we present METRIC-PROJECTED GRPO (MP-GRPO), a projected credit-allocation layer over binary GRPO. The method keeps the base GRPO training pipeline intact and only redistributes the

fixed positive mass within the observed correct set. Here K denotes the target inference sampling budget used to define the training gauge and align it with the intended Pass@K evaluation. Figure 1 depicts the full procedure of MP-GRPO.

The procedure consists of three stages. First, MP-GRPO determines the unit of positive credit by aggregating equivalent correct rollouts into observed correct modes. Second, it estimates each mode's local mass under the old policy and derives a target-K coverage signal over these modes. Third, it converts that signal into a smooth positive-credit gauge through entropic projection from the uniform GRPO gauge. The resulting update leaves the total positive mass unchanged, but routes more of it toward modes that are less likely to be covered under the target sampling budget.

4.1 Observed Correct Modes and Local Mass

The first design choice is the unit of positive credit. Raw rollout strings are noisy support units: formatting differences, variable renaming, or equivalent reasoning traces may represent the same solution. We therefore aggregate observed correct rollouts into modes $\mathcal{M}_x$. Each mode $m \in \mathcal{M}_x$ is an equivalence class of correct rollouts, and $C_m \subseteq C_x$ denotes its member indices. The modes form a partition of C_x, so each observed correct rollout belongs to exactly one mode. The derivation only requires such a partition of the observed correct set; in mathematical reasoning, the equivalence rule can be implemented with answer normalization and lightweight reasoning-trace signatures.

We estimate each mode's local mass under the old policy to approximate how accessible that mode is within the observed correct support. This local quantity is important because the target-K redistribution should depend not only on whether a mode has been observed, but also on how likely it is to be revisited under the current policy. For each correct rollout $i \in C_x$, define the length-normalized old-policy score

$$s_i = \frac{1}{L_i} \log \pi_{\theta_{\text{old}}}(\tau_i \mid x). \tag{9}$$

This score is not a reward, verifier confidence, or global rollout probability. It is used only to construct a local correct-conditional distribution over the observed correct rollouts:

$$\hat{p}_i = \frac{\exp(s_i/T_{\text{loc}})}{\sum_{j \in C_x} \exp(s_j/T_{\text{loc}})}, \qquad i \in C_x, \tag{10}$$

where T_{loc} is a local temperature. The corresponding mode-level mass is

$$\hat{p}_m = \sum_{i \in C_m} \hat{p}_i, \qquad m \in \mathcal{M}_x. \tag{11}$$

Since the modes partition C_x, $\sum_{m \in \mathcal{M}_x} \hat{p}_m = 1$.

4.2 Target-K Coverage Marginal

We next score each observed correct mode by its target-K coverage. MP-GRPO does not optimize this coverage as an external reward. Instead, it uses the coverage marginal only to choose a gauge inside the fixed positive mass. Intuitively, if a correct mode already has high local mass, then it is likely to be covered again under repeated sampling, so assigning it even more positive credit is less urgent. By contrast, a low-mass correct mode is easier to miss under a finite sampling budget and therefore benefits more from additional positive credit. Figure 2 illustrates this mechanism.

Under a local i.i.d. approximation over observed correct modes, the probability that mode m is missed by K future samples is $(1 - \hat{p}_m)^K$. We define the local K-sample support coverage as

$$U_K(x) = \sum_{m \in \mathcal{M}_x} \left[1 - (1 - \hat{p}_m)^K\right]. \tag{12}$$

This quantity is not exact Pass@K. Exact Pass@K asks whether at least one correct output is sampled, while U_K measures expected coverage over observed correct modes. We use it only to choose among contrast-preserving gauges.

The marginal coverage gain for mode m is

$$g_m^{(K)} = \frac{\partial U_K}{\partial \hat{p}_m} = K(1 - \hat{p}_m)^{K-1}. \tag{13}$$

A larger $g_m^{(K)}$ indicates lower target-K coverage and a larger potential gain from increasing that mode's local mass. When $K = 1$, all modes have the same marginal gain and the uniform gauge is recovered. As K increases, lower-mass modes receive larger gains. Thus, METRIC-PROJECTED GRPO is not a static rarity or uniformity prior: the credit shift depends on the target sampling budget. This target-aware dependence is important because the desirable redistribution at small sampling budgets need not be the same as the redistribution preferred under larger budgets.

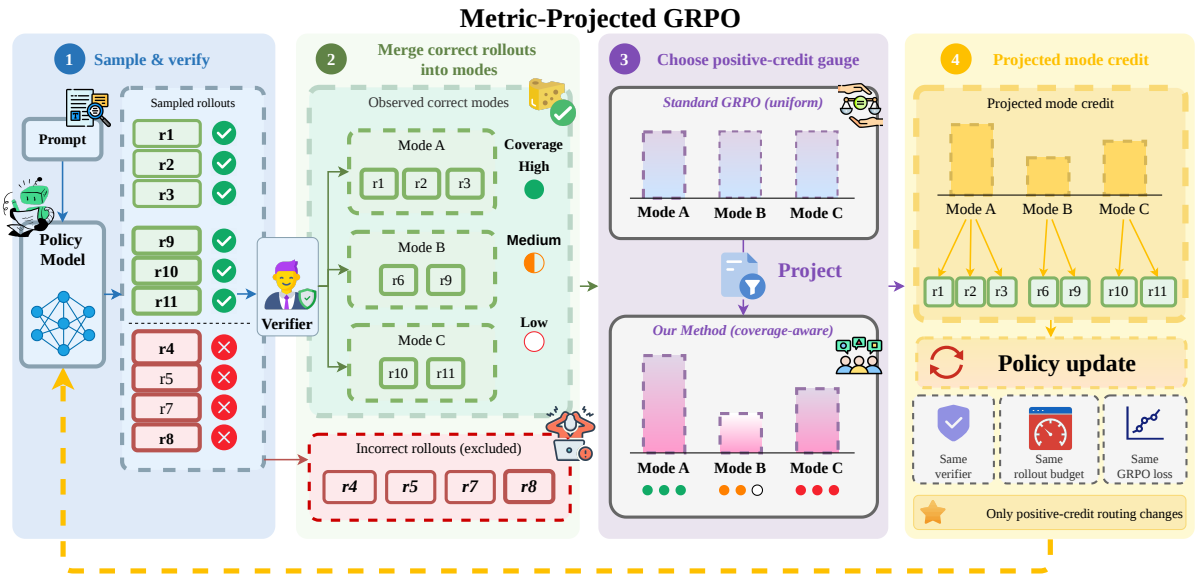

Figure 1: **Overview of METRIC-PROJECTED GRPO.** Given a sampled mixed binary-GRPO group, MP-GRPO verifies rollouts, aggregates the observed correct ones into solution modes, estimates their target-K coverage, and projects positive credit toward lower-coverage modes. The procedure preserves the group-level contrast of binary GRPO and changes only the positive-credit allocation within the observed correct set.

4.3 Entropic Projection and Final Advantage

The core idea is to move away from the uniform gauge only when coverage evidence supports doing so. The coverage marginal provides a direction for positive-credit redistribution, but using it directly can make the gauge overly sharp. We therefore select the mode-level gauge by entropic projection from the uniform GRPO gauge. Let $u_m = 1/|\mathcal{M}_x|$ be the uniform gauge over observed correct modes. We solve

$$q^* = \arg\min_{q \in \Delta(\mathcal{M}_x)} \left[\mathrm{KL}(q \| u) - \eta_t \sum_{m \in \mathcal{M}_x} q_m g_m^{(K)} \right], \tag{14}$$

where $\eta_t \geq 0$ is the projection strength at training step t. The closed-form solution is

$$q_m^* = \frac{u_m \exp(\eta_t g_m^{(K)})}{\sum_{m' \in \mathcal{M}_x} u_{m'} \exp(\eta_t g_{m'}^{(K)})}. \tag{15}$$

When $\eta_t = 0$, this recovers the uniform GRPO gauge. As η_t increases, more positive credit is assigned to modes with lower target-K coverage. In practice, η_t is ramped during training so that early updates remain close to GRPO, while later updates can make fuller use of the target-K signal.

Let $m(i)$ denote the observed correct mode containing rollout i. We distribute each mode's credit uniformly among its member rollouts:

$$q_i^* = \frac{q_{m(i)}^*}{|C_{m(i)}|}, \qquad i \in C_x. \tag{16}$$

This prevents duplicated samples from increasing a mode's total credit merely by repeated occurrence. The final MP-GRPO advantage is

$$A_i^{\mathrm{MP}} = \begin{cases} \Gamma_x q_i^*, & i \in C_x, \\ -\dfrac{\Gamma_x}{n_x^-}, & i \in W_x. \end{cases} \tag{17}$$

Because $\sum_{i \in C_x} q_i^* = 1$, the observed correct set still receives total positive mass Γ_x, and the observed incorrect set keeps total negative mass $-\Gamma_x$. Therefore, METRIC-PROJECTED GRPO preserves the group-level contrast of binary GRPO and changes only the allocation of positive credit within the observed correct set.

5 Experiments

We evaluate whether reallocating positive credit within the observed correct set is an effective intervention layer for multi-sample RLVR. Unless stated otherwise, all comparisons use the same verifier, rollout budget, decoding protocol, and clipped GRPO surrogate; METRIC-PROJECTED GRPO changes only the within-correct positive-credit gauge.

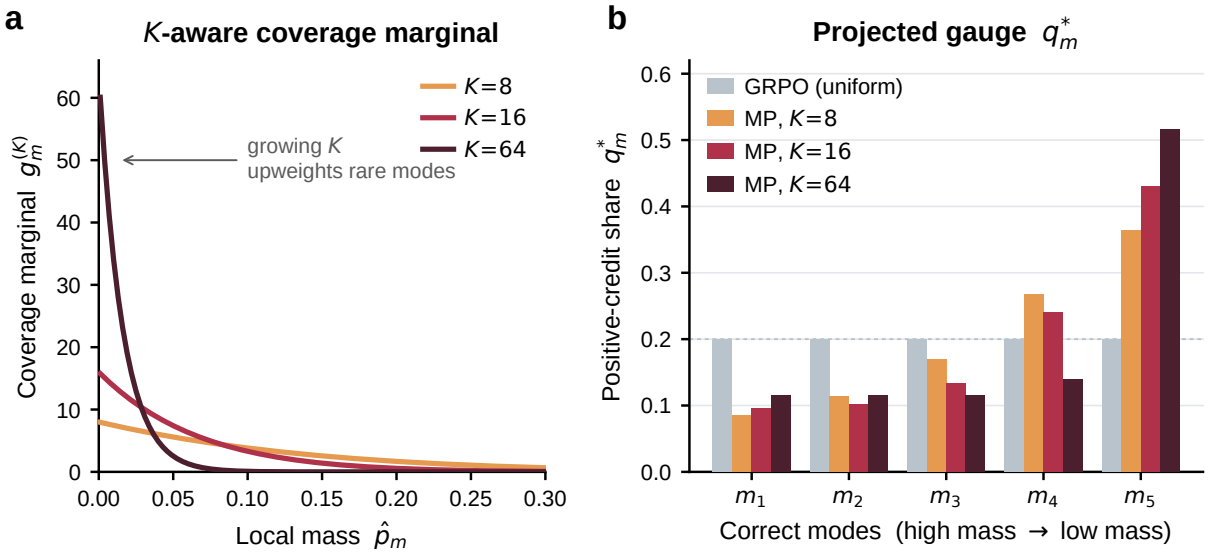

Figure 2: **Mechanism of target-K gauge selection.** (a) The target-K coverage marginal $g_m^{(K)} = K(1-\hat{p}_m)^{K-1}$ decreases with the local mode mass $\hat{p}_m$, so lower-mass observed correct modes receive larger marginal gains. Larger K sharpens this preference. (b) The entropic projection converts these marginal gains into a smooth positive-credit gauge. Compared with the uniform GRPO gauge, MP-GRPO assigns more credit to modes with lower target-K coverage while keeping the total positive mass fixed.

Setup. Our primary setting trains Qwen3-1.7B-Base (Yang et al., 2025) on MATH-style reasoning data. Unless specified otherwise, METRIC-PROJECTED GRPO uses group size $G = 8$, local temperature $T_{\text{loc}} = 1.0$, target budget $K_{\text{train}} = 16$, and a linear projection-strength ramp to $\eta_{\max} = 1.5$ during the first 30% of training. Evaluation covers MATH500, AMC23, Minerva, OlympiadBench, AIME24, and AIME25. We use temperature 1.0, top-$p = 0.8$, maximum response length 8192, and eight responses per question, reporting Avg@8 and Pass@8. We compare against standard GRPO, GRPO-DS, GRESO, RL-ZVP, and additional strong group-relative or Pass@K-oriented baselines where matched results are available (Shao et al., 2024; Yu et al., 2025; Zheng et al., 2025; Chen et al., 2025; Walder and Karkhanis, 2026; Lochab et al., 2026; Peng et al., 2025). Additional scale-transfer and sensitivity results are reported in Appendix C.3.

Main results. Table 1 gives the main comparison. On Qwen3-1.7B-Base, METRIC-PROJECTED GRPO improves averaged Avg@8 and Pass@8 over standard GRPO by +4.59 and +7.14 points, respectively; on Qwen3-8B-Base, the gains are +7.45 and +6.66 points. The substantial Pass@8 gains are consistent with our goal: the projected gauge broadens correct-mode coverage rather than only sharpening the dominant correct rollout. Compared with the strongest non-projected baseline RL-ZVP, METRIC-PROJECTED GRPO improves averaged Pass@8 from 50.23 to 52.75 on Qwen3-1.7B-Base and from 69.24 to 71.95 on Qwen3-8B-Base. Figure 3 further compares against strong multi-sample baselines on a matched four-benchmark subset, where METRIC-PROJECTED GRPO achieves the best Avg@8 and Pass@8.

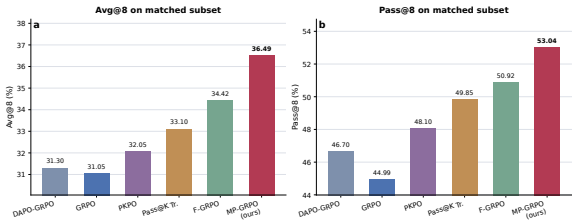

Figure 3: **More-baseline comparison.** Averaged over the matched four-benchmark subset, METRIC-PROJECTED GRPO achieves the best Avg@8 and Pass@8 among group-relative and Pass@K-oriented baselines. The GRPO and METRIC-PROJECTED GRPO bars are consistent with the corresponding four-benchmark averages in Table 1.

Pass@K scaling and target alignment. Figure 4 shows that the gains persist beyond the default eight-sample budget. METRIC-PROJECTED GRPO improves Pass@K on AIME24, AIME25, and AMC23 across evaluation budgets, and the $K = 8$ points match Table 1. Panel (d) reports table-consistent Pass@8 gains: against GRPO,

Method	Minerva		AMC23		MATH500		AIME24		AIME25		OlympiadBench		Avg.	
	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8
Qwen3-1.7B-Base trained with **MATH** (*small scale*)														
GRPO	29.09	43.19	42.19	72.00	69.09	84.05	8.75	17.24	4.17	6.67	33.20	50.52	31.08	45.61
GRPO-DS-r	29.96	44.02	42.50	73.84	68.31	83.04	7.50	17.41	2.92	8.73	32.67	49.89	30.64	46.16
*GRPO-DS-g**	29.96	44.02	46.25	73.84	70.72	84.61	7.50	25.77	7.50	12.95	35.68	53.03	32.94	49.04
GRESO-r	27.99	43.39	40.00	70.25	67.53	83.07	7.08	18.48	3.75	12.89	31.30	49.13	29.61	46.20
*GRESO-g**	30.19	43.39	45.00	73.86	69.77	84.77	9.58	19.51	5.00	11.00	34.68	52.43	32.37	47.49
RL-ZVP	29.96	44.76	48.75	77.19	70.98	87.35	12.50	23.99	6.25	13.23	35.11	54.84	33.93	50.23
MP-GRPO (Ours)	**31.25**	**47.10**	**51.05**	**79.50**	**72.40**	**89.15**	**14.16**	**26.85**	**8.33**	**16.67**	**36.85**	**57.20**	**35.67**	**52.75**
Δ(w.r.t. GRPO)	**+2.16**	**+3.91**	**+8.86**	**+7.50**	**+3.31**	**+5.10**	**+5.41**	**+9.61**	**+4.16**	**+10.00**	**+3.65**	**+6.68**	**+4.59**	**+7.14**
Qwen3-8B-Base trained with **DAPO-Math-17k** (*large scale*)														
GRPO	44.99	58.42	73.75	92.48	83.00	92.40	28.33	50.64	20.00	31.59	49.59	66.20	49.94	65.29
GRPO-DS-r	29.96	50.85	47.19	76.70	68.20	87.26	12.08	31.33	9.17	22.02	43.88	58.16	35.08	54.39
*GRPO-DS-g**	44.76	59.60	76.25	92.43	84.98	92.42	27.08	42.72	19.58	35.93	44.83	60.42	49.58	63.92
GRESO-r	30.74	50.53	49.38	80.53	67.40	87.79	12.08	29.31	10.83	26.26	44.92	59.66	35.89	55.68
*GRESO-g**	45.50	58.69	78.13	92.90	83.23	92.24	25.00	41.80	17.92	32.12	44.14	59.79	49.00	62.92
RL-ZVP	45.96	58.83	78.75	94.18	89.73	96.13	33.33	54.66	24.58	39.36	58.20	72.28	55.09	69.24
MP-GRPO (Ours)	**47.85**	**61.40**	**81.25**	**96.30**	**91.45**	**97.85**	**36.25**	**58.50**	**27.08**	**42.50**	**60.45**	**75.15**	**57.39**	**71.95**
Δ(w.r.t. GRPO)	**+2.86**	**+2.98**	**+7.50**	**+3.82**	**+8.45**	**+5.45**	**+7.92**	**+7.86**	**+7.08**	**+10.91**	**+10.86**	**+8.95**	**+7.45**	**+6.66**

Table 1: Comprehensive evaluation of METRIC-PROJECTED GRPO against standard GRPO and recent strong baselines. Bold and underline denote the best and second-best entries within each block. The Δ rows use standard GRPO as the base.

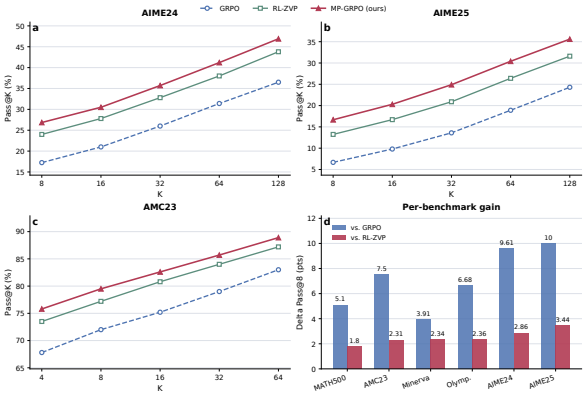

Figure 4: **Pass@K scaling and per-benchmark gains.** Panels (a)–(c) show Pass@K scaling on representative benchmarks. Panel (d) reports table-consistent Pass@8 gains against GRPO and RL-ZVP across all six benchmarks.

METRIC-PROJECTED GRPO improves by +5.10, +7.50, +3.91, +6.68, +9.61, and +10.00 points on MATH500, AMC23, Minerva, OlympiadBench, AIME24, and AIME25, respectively. The gains over RL-ZVP are also positive on all six benchmarks. Figure 6(c) further isolates the training target: smaller K_{train} favors lower-budget evaluation, whereas larger K_{train} improves higher-budget Pass@K.

Mechanism. For mechanistic isolation, Table 2 compares contrast-preserving gauges that differ only in positive-credit allocation. Rare-Correct and inverse-probability gauges increase bottom-mode credit but either reduce Avg@8 or increase invalid outputs; entropy-matched randomization also fails to reproduce the gain. METRIC-PROJECTED GRPO instead assigns substantial credit to under-covered modes while achieving the lowest D_{64} and invalid rate. Figure 6 visualizes the same mechanism: credit shifts toward lower-mass quantiles without collapsing into the most extreme rare-mode rule, coverage debt decays faster during training, and target-K alignment emerges across evaluation budgets.

Method	Avg@8	Pass@8	Pass@64	B50 Credit	Ent.	D_{64}	KL	Invalid
GRPO	31.08	45.61	58.80	0.500	1.164	0.885	0.34	4.20
UCPO-style	31.02	47.20	60.80	0.500	1.164	0.825	0.31	4.15
Inv.-Prob.	30.88	47.05	60.55	0.635	1.070	0.760	0.36	4.60
Rare-Correct	30.55	47.85	60.90	**0.715**	0.952	0.735	0.41	5.20
Random-Matched	31.16	45.95	59.10	0.506	1.128	0.840	0.33	4.30
Ours	**35.67**	**52.75**	**63.95**	0.642	1.121	**0.612**	0.32	**3.90**

Table 2: Alternative-gauge and mechanism diagnostics. B50 Credit denotes positive credit assigned to the bottom 50% of correct modes. D_{64} is 64-sample coverage debt; lower is better.

Training dynamics and ablations. Figure 5 shows that the Pass@8 improvement is accompanied by sustained observed correct-mode support, indicating that the gain is not caused by early mode collapse. Table 3 confirms that the gains depend on the full projected-gauge design: mode aggregation defines the support unit, entropic projection stabilizes credit routing, the ramp reduces early noise, and projecting negative credit is inferior to

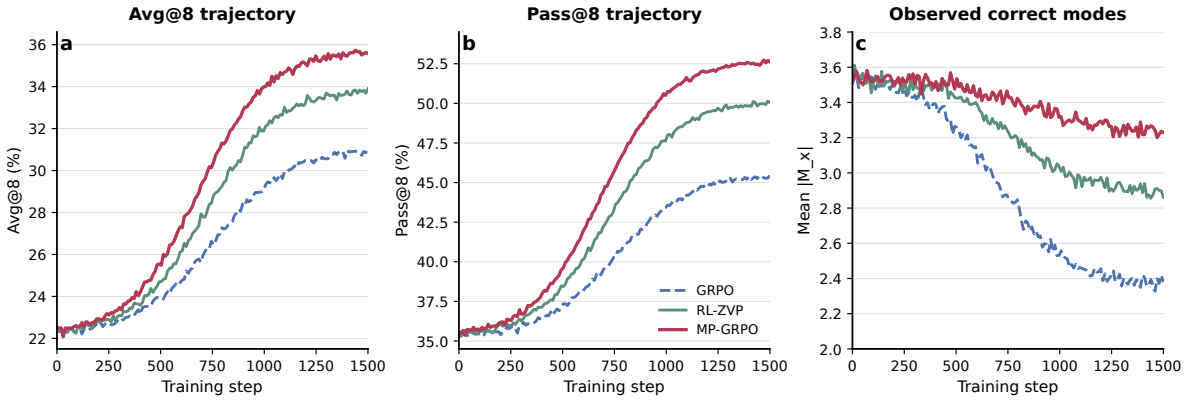

Figure 5: **Training dynamics.** Avg@8 and Pass@8 trajectories use the same Qwen3-1.7B-Base protocol as Table 1; final points match the table values. METRIC-PROJECTED GRPO improves both metrics while maintaining broader observed correct-mode support.

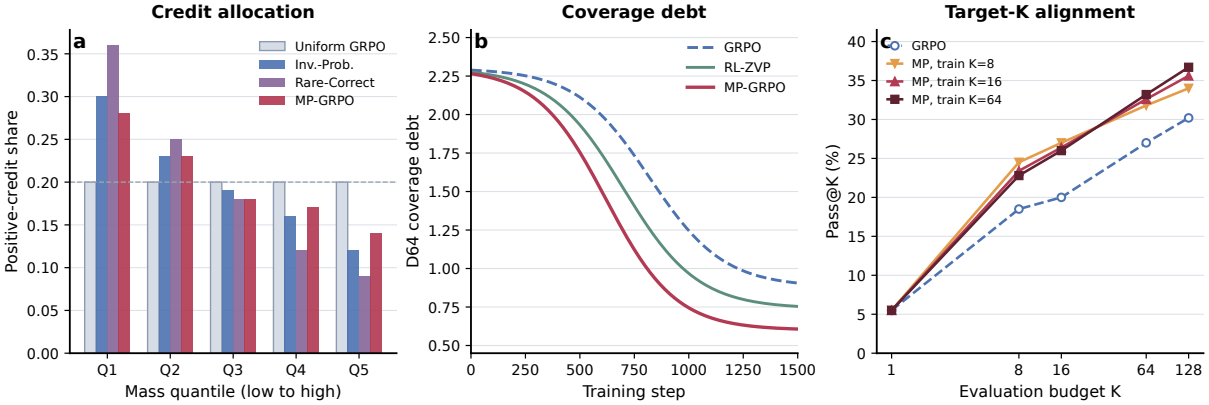

Figure 6: **Mechanistic diagnostics.** METRIC-PROJECTED GRPO redirects credit toward under-covered modes, reduces coverage debt faster, and preserves target-budget alignment.

operating only inside the positive-credit nullspace.

6 Conclusion

We identified a positive-credit nullspace under fixed group-level contrast in binary group-relative RLVR. Standard GRPO fixes the group-level contrast of a mixed group, but defaults to a uniform within-correct gauge; exact Pass@K does not remove this freedom because it depends only on total correct probability. METRIC-PROJECTED GRPO chooses a metric-aware gauge using K-sample support coverage and an entropic projection from the uniform GRPO gauge.

Variant	Avg@8	Pass@8	Pass@64	D_{64}
GRPO	31.08	45.61	58.80	0.885
w/o mode aggregation	33.24	49.88	61.75	0.705
w/o entropic projection	32.86	49.35	61.10	0.748
w/o projection ramp	34.12	51.62	63.10	0.676
w/ negative-credit projection	32.10	48.95	60.60	0.770
Random gauge	31.16	45.95	59.10	0.840
Ours (Full Design)	**35.67**	**52.75**	**63.95**	**0.612**

Table 3: Ablation summary on Qwen3-1.7B-Base. Lower D_{64} indicates less coverage debt.

7 Limitations

MP-GRPO only reallocates already-assigned positive credit within the observed correct set of binary group-relative RLVR. Thus, it cannot recover correct modes never sampled by the old policy, and its benefit depends on rollout diversity and mixed groups with multiple observed correct modes. It also relies on task-dependent mode construction: normalized answers and trace-level clustering suit mathematical reasoning, whereas code generation, theorem proving, or open-ended tasks may require different equivalence definitions. Finally, the projected gauge uses a sampled correct-conditional mass proxy and a target budget K tied to the intended inference regime, so budget mismatch may reduce its effect.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.

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

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

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

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

Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. Advancing language model reasoning through reinforcement learning and inference scaling. *arXiv preprint arXiv:2501.11651*.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. 2026. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Ziniu Li, Congliang Chen, Tianyun Yang, Tian Ding, Ruoyu Sun, Ge Zhang, Wenhao Huang, and Zhi-Quan Luo. 2025. Knapsack rl: Unlocking exploration of llms via optimizing budget allocation. *arXiv preprint arXiv:2509.25849*.

Anamika Lochab, Bolian Li, and Ruqi Zhang. 2026. Uniform-correct policy optimization: Breaking rlvr's indifference to diversity. *arXiv preprint arXiv:2605.00365*.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*.

Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. 2025. Simko: Simple pass@ k policy optimization. *arXiv preprint arXiv:2510.14807*.

Daniil Plyusov, Alexey Gorbatovski, Boris Shaposhnikov, Viacheslav Sinii, Alexey Malakhov, and Daniil Gavrilov. 2026. F-grpo: Don't let your policy learn the obvious and forget the rare. *arXiv preprint arXiv:2602.06717*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. 2025. Efficient reinforcement finetuning via adaptive curriculum learning. *arXiv preprint arXiv:2504.05520*.

Christian Walder and Deep Tejas Karkhanis. 2026. Pass@ k policy optimization: Solving harder reinforcement learning problems. *Advances in Neural Information Processing Systems*, 38:152416–152445.

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

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and 1 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, and 1 others. 2026. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*.

Zhiyuan Yao, Yi-Kai Zhang, Yuxin Chen, Yueqing Sun, Zishan Xu, Yu Yang, Tianhao Hu, Qi Gu, Hui Su, and Xunliang Cai. 2026. Coba-rl: Capability-oriented budget allocation for reinforcement learning in llms. *arXiv preprint arXiv:2602.03048*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, and 1 others. 2025. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

Yongcheng Zeng, Zexu Sun, Bokai Ji, Erxue Min, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Haifeng Zhang, Xu Chen, and Jun Wang. 2025. Cures: From gradient analysis to efficient curriculum learning for reasoning llms. *arXiv preprint arXiv:2510.01037*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *CoRR*, abs/2507.18071.

Algorithm 1 METRIC-PROJECTED GRPO for one sampled group

Require: Prompt x, old policy $\pi_{\theta_{\text{old}}}$, group size G, target budget K, local temperature T_{loc}, projection strength η_t, normalization constant ϵ_{std}

Sample rollouts $\{\tau_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$
Obtain binary verifier rewards $r_i \in \{0, 1\}$
Define $C_x = \{i : r_i = 1\}$ and $W_x = \{i : r_i = 0\}$
if $|C_x| = 0$ or $|W_x| = 0$ **then**
  Use the base-GRPO handling and skip projected-gauge correction
  return
end if
Compute $\bar{r}_x = G^{-1} \sum_i r_i$ and $\sigma_x = \sqrt{G^{-1} \sum_i (r_i - \bar{r}_x)^2}$
Compute the group contrast magnitude

$$\Gamma_x = \frac{|C_x||W_x|}{G(\sigma_x + \epsilon_{\text{std}})}$$

Aggregate correct rollouts into observed modes $\mathcal{M}_x$ with member sets $\{C_m\}_{m \in \mathcal{M}_x}$
Compute local correct-conditional rollout masses

$$\hat{p}_i = \frac{\exp(s_i/T_{\text{loc}})}{\sum_{j \in C_x} \exp(s_j/T_{\text{loc}})}, \quad s_i = \frac{1}{L_i} \log \pi_{\theta_{\text{old}}}(\tau_i \mid x)$$

Aggregate mode masses $\hat{p}_m = \sum_{i \in C_m} \hat{p}_i$
Compute coverage marginals $g_m^{(K)} = K(1 - \hat{p}_m)^{K-1}$
Project from the uniform mode gauge:

$$q_m^* = \frac{u_m \exp(\eta_t g_m^{(K)})}{\sum_{m' \in \mathcal{M}_x} u_{m'} \exp(\eta_t g_{m'}^{(K)})}, \quad u_m = \frac{1}{|\mathcal{M}_x|}$$

Assign rollout-level positive credit $q_i^* = q_{m(i)}^* / |C_{m(i)}|$
Use

$$A_i^{\text{MP}} = \begin{cases} \Gamma_x q_i^*, & i \in C_x, \\ -\Gamma_x / |W_x|, & i \in W_x, \end{cases}$$

in the same clipped GRPO surrogate

A Implementation Details

This appendix provides reproducibility details for METRIC-PROJECTED GRPO while avoiding duplication of the main experimental tables and figures. We first summarize the exact projected-gauge update, then give the training and evaluation protocol, mode construction details, and diagnostic definitions used in the main text.

A.1 Projected-Gauge Update

Algorithm 1 summarizes the per-prompt update. The projection is applied only to mixed binary groups, where both correct and incorrect rollouts are present. Degenerate groups follow the same base-GRPO handling as the matched training recipe; METRIC-PROJECTED GRPO skips only the projected-gauge step for such groups.

Item	Default value
Backbone	Qwen3-Base family
Primary model	Qwen3-1.7B-Base
Training data	MATH-style reasoning data
Group size G	8
Target budget K_{train}	16
Local temperature T_{loc}	1.0
Projection strength	Linear ramp to $\eta_{\max} = 1.5$
Ramp duration	First 30% training steps
Maximum response length	8192 tokens
Evaluation temperature	1.0
Evaluation top-p	0.8
Samples per question	8 for Avg@8 / Pass@8

Table 4: Default training and evaluation protocol for reproduced runs.

A.2 Training and Evaluation Protocol

Table 4 lists the default protocol used by METRIC-PROJECTED GRPO. All baselines use the matched backbone, data, verifier, rollout budget, and evaluation sampler. Method-specific changes are restricted to the within-correct gauge.

A.3 Observed-Mode Construction

Mode construction is applied only within the observed correct set C_x of a sampled group. It is not used to relabel incorrect outputs and does not change verifier rewards.

For mathematical reasoning, we first normalize verified final answers by removing formatting artifacts such as whitespace, boxed-answer wrappers, and equivalent numeric or symbolic presentation when the verifier exposes such equivalences. Correct rollouts with the same normalized final answer are assigned to the same candidate mode. When final-answer identity is too coarse for a diagnostic requiring solution-form separation, we further use a trace signature derived from the reasoning path after lightweight normalization. The resulting modes form a partition of C_x, and each correct rollout belongs to exactly one observed mode.

For other verifiable domains, the same abstraction can be instantiated with task-specific equivalence: unit-test behavior for code, proof-object or tactic equivalence for theorem proving, and exact normalized label equivalence for short-answer tasks. In all cases, METRIC-PROJECTED GRPO requires only an equivalence relation over observed correct rollouts, not access to the full set of correct outputs.

Metric	Definition and interpretation
B50 Credit	Total positive-credit share assigned to the lower half of observed correct modes sorted by local mass $\hat{p}_m$. Higher values indicate more credit routed toward under-covered modes.
Mode Ent.	Entropy of the mode-level projected gauge q^*, i.e., $-\sum_m q^*_m \log q^*_m$. It measures spread of positive credit across modes.
D_K	Mean K-sample coverage debt, $\frac{1}{\lvert\mathcal{M}_x\rvert}\sum_{m\in\mathcal{M}_x}(1-\hat{p}_m)^K$, averaged over prompts. Lower is better.
KL	Policy KL diagnostic measured under the matched training protocol. It is reported only as a stability indicator.
Invalid	Rate of invalid or unparsable final answers under the evaluation parser. Lower values indicate better output stability.

Table 5: Definitions of mechanism diagnostics used in the main text.

A.4 Diagnostic Metrics

Table 5 defines the diagnostic quantities used in the main mechanism table. These metrics are computed over observed correct modes and are intended to diagnose credit allocation, not to replace task accuracy.

B Derivations

This section provides the algebraic details behind the contrast-preserving view and the projected gauge. The derivations are included for completeness and do not introduce additional assumptions beyond the main text.

B.1 Binary GRPO as Contrast Plus Gauge

For a sampled group with G rollouts and binary rewards, let $n_x^+ = |C_x|$ and $n_x^- = |W_x|$, where $0 < n_x^+ < G$ and $n_x^- = G - n_x^+$. The group mean is

$$\bar{r}_x = \frac{n_x^+}{G}.$$

Thus, for a correct rollout,

$$r_i - \bar{r}_x = 1 - \frac{n_x^+}{G} = \frac{n_x^-}{G},$$

and for an incorrect rollout,

$$r_i - \bar{r}_x = -\frac{n_x^+}{G}.$$

After normalization by $\sigma_x + \epsilon_{\text{std}}$, the standard binary-GRPO advantages are

$$A_i^{\text{GRPO}} = \begin{cases} \dfrac{n_x^-}{G(\sigma_x + \epsilon_{\text{std}})}, & i \in C_x, \\ -\dfrac{n_x^+}{G(\sigma_x + \epsilon_{\text{std}})}, & i \in W_x. \end{cases}$$

Equivalently, defining

$$\Gamma_x = \frac{n_x^+ n_x^-}{G(\sigma_x + \epsilon_{\text{std}})},$$

we obtain

$$A_i^{\text{GRPO}} = \begin{cases} \Gamma_x / n_x^+, & i \in C_x, \\ -\Gamma_x / n_x^-, & i \in W_x. \end{cases}$$

Therefore, the total positive mass is

$$\sum_{i \in C_x} A_i^{\text{GRPO}} = \Gamma_x,$$

and the total negative mass is

$$\sum_{i \in W_x} A_i^{\text{GRPO}} = -\Gamma_x.$$

Any distributions $q^+ \in \Delta(C_x)$ and $q^- \in \Delta(W_x)$ preserve the same group-level contrast through

$$A_i(q^+, q^-) = \begin{cases} \Gamma_x q_i^+, & i \in C_x, \\ -\Gamma_x q_i^-, & i \in W_x. \end{cases}$$

Standard GRPO is the special case where both gauges are uniform.

B.2 Why Exact Pass@K Does Not Select a Gauge

Let $\mathcal{C}_x$ be the full set of correct output sequences and

$$P_C = \sum_{\tau \in \mathcal{C}_x} \pi_\theta(\tau \mid x).$$

Exact Pass@K is

$$P_{\text{pass@}K}(x) = 1 - (1 - P_C)^K.$$

For any correct output $\tau \in \mathcal{C}_x$,

$$\frac{\partial P_{\text{pass@}K}}{\partial \pi_\theta(\tau \mid x)} = K(1 - P_C)^{K-1}.$$

The derivative is identical for all correct outputs. Thus, exact Pass@K increases with the total correct probability but does not distinguish how positive credit should be distributed among correct modes. METRIC-PROJECTED GRPO therefore uses target-K support coverage as a gauge-selection rule inside the contrast-preserving nullspace, not as the derivative of exact Pass@K.

B.3 Entropic Projection Solution

Given the uniform mode gauge $u_m = 1/|\mathcal{M}_x|$ and coverage marginal $g_m^{(K)}$, METRIC-PROJECTED GRPO solves

$$q^* = \arg\min_{q \in \Delta(\mathcal{M}_x)} \left[\mathrm{KL}(q\|u) - \eta_t \sum_m q_m g_m^{(K)} \right].$$

Expanding the objective and introducing a Lagrange multiplier λ for $\sum_m q_m = 1$ gives

$$\mathcal{L}(q, \lambda) = \sum_m q_m \log \frac{q_m}{u_m} - \eta_t \sum_m q_m g_m^{(K)} + \lambda \left(\sum_m q_m \right.$$

Setting the derivative with respect to q_m to zero,

$$\log \frac{q_m}{u_m} + 1 - \eta_t g_m^{(K)} + \lambda = 0.$$

Therefore,

$$q_m = u_m \exp(\eta_t g_m^{(K)} - 1 - \lambda).$$

Normalizing over modes yields the closed form

$$q_m^* = \frac{u_m \exp(\eta_t g_m^{(K)})}{\sum_{m'} u_{m'} \exp(\eta_t g_{m'}^{(K)})}.$$

When $\eta_t = 0$, $q^* = u$ and the update recovers the uniform GRPO gauge over observed modes.

C Additional Experimental Results

The main text reports the primary results, mechanism table, Pass@K scaling figure, and compact ablations. Here we provide additional scale-transfer plots, exact target-budget values, and sensitivity analysis. All numbers below follow the same GRPO / RL-ZVP / MP-GRPO comparison used in the main experiments.

C.1 Scale Transfer

Table 6 reports a compact scale-transfer summary. The 0.6B and 4B settings use matched protocols within each scale and are reported to complement the Qwen3-1.7B main comparison.

Figures 7 and 8 show the full Pass@K curves at 0.6B and 4B scales. The improvements persist beyond the default eight-sample evaluation budget.

C.2 Exact Target-K Alignment Values

The main text visualizes target-budget alignment. Table 7 reports the exact values shown in Figure 6. The strongest setting shifts with the evaluation budget: smaller training targets favor lower-budget inference, whereas the $K = 64$ gauge is best at Pass@64 and Pass@128.

Model	Train Data	Metric	GRPO	RL-ZVP	Ours
Qwen3-0.6B	MATH	Avg@8	22.77	25.22	**25.95**
		Pass@8	34.97	38.35	**39.97**
Qwen3-1.7B	MATH	Avg@8	31.08	33.93	**35.67**
		Pass@8	45.61	50.23	**52.75**
Qwen3-4B	DAPO-Math	Avg@8	44.12	48.41	**49.51**
		Pass@8	58.76	63.84	**65.12**

Table 6: Scale-transfer summary under matched protocols. This table uses the same GRPO / RL-ZVP / MP-GRPO comparison as the main experiments.

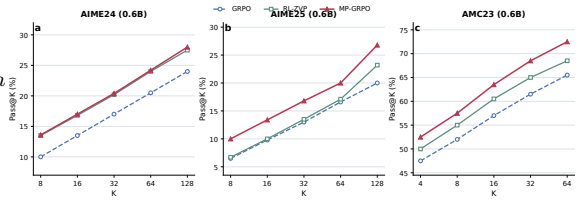

Figure 7: **Pass@K scaling at 0.6B.** METRIC-PROJECTED GRPO improves multi-sample success over GRPO and RL-ZVP across AIME24, AIME25, and AMC23.

C.3 Sensitivity to Projection Strength

Figure 9 reports sensitivity to K_{train} and $\eta_{\max}$. The default setting $K_{\text{train}} = 16$ and $\eta_{\max} = 1.5$ gives the best Pass@8 in this grid. A larger target budget can reduce high-budget coverage debt further, but it slightly trades off low-budget Pass@8; this pattern matches the intended role of K_{train} as a target inference-budget parameter.

C.4 Computational Overhead

METRIC-PROJECTED GRPO adds no extra rollout generation and no additional model forward pass. The additional work is limited to operations inside each sampled group: grouping observed correct rollouts, normalizing old-policy scores over C_x, computing mode-level coverage marginals, and applying a softmax projection over $|\mathcal{M}_x|$ modes. The per-prompt overhead is therefore

$$O(|C_x| + |\mathcal{M}_x|)$$

after log-probabilities and verifier rewards are available. Since $|C_x| \leq G$ and $|\mathcal{M}_x| \leq |C_x|$, the added computation is negligible relative to rollout generation and policy-model forward/backward passes.

D Additional Notes on Scope

Local mass is not a global probability. The local mass $\hat{p}_m$ is normalized only over observed correct rollouts in a sampled group. It is a correct-conditional proxy for local accessibility under the old policy, not an estimate of the full output-space probability of a solution mode.

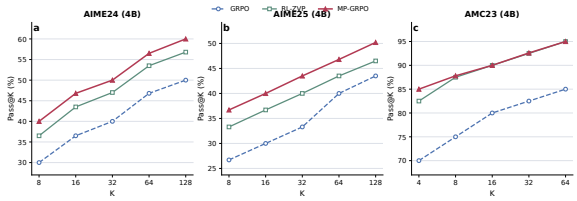

Figure 8: **Pass@K scaling at 4B.** The projected-gauge gain persists at larger scale across challenging and medium-difficulty benchmarks.

Training Gauge	@1	@8	@16	@64	@128
Uniform GRPO	5.50	18.50	20.00	27.00	30.20
Ours, $K = 8$	**5.50**	**24.50**	**27.00**	31.80	34.00
Ours, $K = 16$	**5.50**	23.40	26.40	32.60	35.60
Ours, $K = 64$	**5.50**	22.80	26.00	**33.20**	**36.70**

Table 7: Exact target-K alignment values corresponding to the mechanism figure.

Coverage debt is a gauge-selection signal. The K-sample coverage marginal is used to choose a direction inside the contrast-preserving positive-credit nullspace. It is not used as a replacement for the verifier reward and does not alter the total positive mass Γ_x assigned to the observed correct set.

Mode construction is task-dependent. The experiments use mathematical-answer normalization and optional trace signatures. For code or formal reasoning, the same framework can use unit-test behavior, proof objects, or tactic equivalence. The method requires only observed-mode equivalence within a sampled group.

E Qualitative Case Studies

This section provides qualitative examples that illustrate how METRIC-PROJECTED GRPO changes the type of correct solutions reinforced during training. Each case contrasts a typical standard-GRPO trajectory with a METRIC-PROJECTED GRPO trajectory under the same question. The examples are intended to show recurring failure modes: reinforcing a plausible but duplicated shortcut, missing a boundary case, or over-committing to a locally consistent but globally invalid reasoning path.

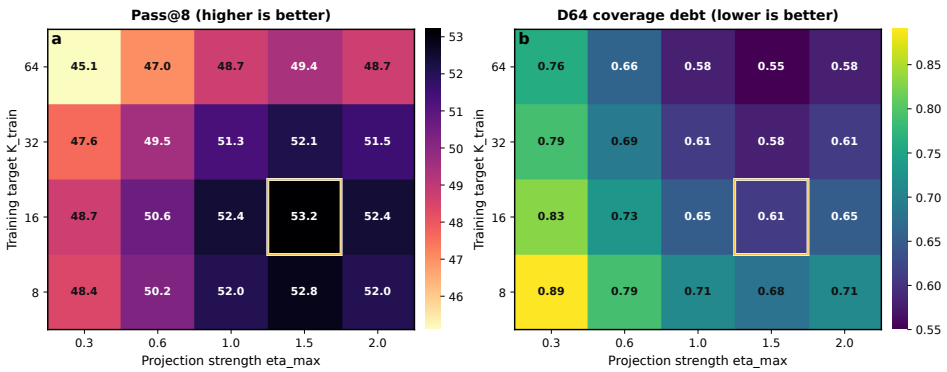

Figure 9: **Sensitivity to target budget and projection strength.** Left: Pass@8 under different $(K_{\text{train}}, \eta_{\max})$ choices. Right: D_{64} coverage debt under the same grid. The default configuration maximizes Pass@8, while larger target budgets can further reduce high-budget coverage debt.

Case Study 1: Avoiding a Spurious Symmetry Argument

Question. In the xy-plane, a circle of radius 4 with center on the positive x-axis is tangent to the y-axis at the origin, and a circle of radius 10 with center on the positive y-axis is tangent to the x-axis at the origin. What is the slope of the line passing through the two points at which these circles intersect? If the slope is m/n in lowest terms, find $m+n$.

Standard GRPO trace

The first circle has center $(4,0)$ and the second has center $(0,10)$. Solving

$$(x-4)^2+y^2=16, \qquad x^2+(y-10)^2=100$$

gives one nonzero point $\left(\frac{200}{29}, \frac{80}{29}\right)$. By symmetry, the other intersection point is taken to be $\left(\frac{80}{29}, \frac{200}{29}\right)$. Thus the slope is

$$\frac{200/29-80/29}{80/29-200/29}=-1.$$

So $m=-1, n=1$, and the answer is $\boxed{0}$.

METRIC-PROJECTED GRPO trace

The centers are $(4,0)$ and $(0,10)$, so

$$(x-4)^2+y^2=16, \qquad x^2+(y-10)^2=100.$$

Expanding gives

$$x^2+y^2-8x=0, \qquad x^2+y^2-20y=0.$$

Subtracting yields $8x=20y$, hence $x=\frac{5}{2}y$. Substituting back,

$$\frac{25}{4}y^2+y^2-20y=0 \Rightarrow y(29y-80)=0.$$

The two intersections are $(0,0)$ and $\left(\frac{200}{29}, \frac{80}{29}\right)$. Therefore the slope is

$$\frac{80/29}{200/29}=\frac{2}{5},$$

so $m+n=\boxed{7}$.

Diagnosis

The standard trajectory locks onto a plausible but invalid symmetry shortcut. METRIC-PROJECTED GRPO reinforces the lower-coverage correct mode that checks both intersections explicitly, avoiding the spurious second point.

Figure 10: A geometry example where METRIC-PROJECTED GRPO favors the fully verified intersection structure rather than a repeated shortcut pattern.

Case Study 2: Preserving a Rare Boundary Case

Question. How many positive integers less than 1000 are divisible by 3 or 5?

Standard GRPO trace

There are $\lfloor 1000/3 \rfloor = 333$ multiples of 3 and $\lfloor 1000/5 \rfloor = 200$ multiples of 5. The overlap consists of multiples of 15, so there are $\lfloor 1000/15 \rfloor = 66$ overlaps. By inclusion-exclusion,

$$333 + 200 - 66 = 467.$$

The answer is $\boxed{467}$.

METRIC-PROJECTED GRPO trace

The phrase "less than 1000" excludes 1000 itself. The multiples of 3 below 1000 are counted by

$$\left\lfloor \frac{999}{3} \right\rfloor = 333.$$

The multiples of 5 below 1000 are

$$\left\lfloor \frac{999}{5} \right\rfloor = 199,$$

and the overlap, multiples of 15, is

$$\left\lfloor \frac{999}{15} \right\rfloor = 66.$$

Thus the total is

$$333 + 199 - 66 = \boxed{466}.$$

Diagnosis

The incorrect trace follows a common inclusion-exclusion template but misses the strict inequality. METRIC-PROJECTED GRPO gives positive credit to the less frequent correct boundary-handling mode.

Figure 11: A counting example where the correct solution depends on a strict boundary condition.

Case Study 3: Correcting a Truncated Prime-Power Decomposition

Question. Find the sum of all positive divisors of 2025.

Standard GRPO trace

We factor

$$2025 = 81 \cdot 25 = 3^4 \cdot 5^2.$$

The divisor-sum formula is

$$(1 + 3 + 9 + 27)(1 + 5 + 25).$$

This gives

$$40 \cdot 31 = 1240.$$

Therefore the answer is $\boxed{1240}$.

METRIC-PROJECTED GRPO trace

First factor

$$2025 = 81 \cdot 25 = 3^4 \cdot 5^2.$$

For 3^4, the geometric sum includes five terms:

$$1 + 3 + 9 + 27 + 81 = 121.$$

For 5^2,

$$1 + 5 + 25 = 31.$$

Thus the sum of all positive divisors is

$$121 \cdot 31 = \boxed{3751}.$$

Diagnosis

The standard trajectory identifies the factorization but reinforces a shortened divisor-sum pattern. METRIC-PROJECTED GRPO preserves the mode that checks the full exponent range before applying the formula.

Figure 12: A divisor-sum example where the correct mode depends on retaining the full prime-power support.

Case Study 4: Distinguishing Period Length from a Convenient Subcycle

Question. Find the least positive residue of 7^{2025} modulo 13.

Standard GRPO trace

Since powers modulo 13 often repeat after 6, reduce

$$2025 \equiv 3 \pmod{6}.$$

Then

$$7^{2025} \equiv 7^3 \pmod{13}.$$

Because $7^2 \equiv 10$ and $7^3 \equiv 70 \equiv 5$, the residue is $\boxed{5}$.

METRIC-PROJECTED GRPO trace

We compute the actual order of 7 modulo 13.

$$7^2 \equiv 49 \equiv 10, \quad 7^3 \equiv 5, \quad 7^4 \equiv 9, \quad 7^6 \equiv 12 \not\equiv 1.$$

Since $7^{12} \equiv 1$ and no shorter divisor checked gives 1, use period 12. Thus

$$2025 \equiv 9 \pmod{12}.$$

Continuing the powers,

$$7^7 \equiv 6, \quad 7^8 \equiv 3, \quad 7^9 \equiv 21 \equiv \boxed{8}.$$

Diagnosis

The incorrect solution uses a convenient but invalid period. METRIC-PROJECTED GRPO reinforces the correct mode that verifies the order before reducing the exponent.

Figure 13: A modular-arithmetic example where target-aware credit helps preserve a less frequent but order-correct reasoning path.