HIGHER-ORDER EXPLORATION IN RLOO WITH PAIRWISE-INDEPENDENT ROLLOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse rewards can leave an entire rollout group without a learning signal. We introduce HO-RLOO, an autoregressive sampler that induces higher-order dependence while preserving the independent law of every trajectory pair and therefore the expected raw RLOO gradient under ideal continuous randomness. A calibrated mixture of distinct and shared probability strata changes interactions among three and four trajectories without changing pairwise laws. In an aligned sparse-reward model, HO-RLOO reduces gradient variance and achieves maximal group coverage among pairwise-independent samplers. On mathematics and code, HO-RLOO improves domain-balanced pass@1 over iid-RLOO by 2.51 and 2.78 points at two Qwen3 model scales under a common training-token cap. Frozen-policy diagnostics show increased mixed-reward groups and rollout diversity on the evaluated prompt panel. These results demonstrate that preserving pairwise independence provides a principled way to shape higher-order exploration while retaining the original RLOO objective.

1 INTRODUCTION

Group-based reinforcement fine-tuning uses several responses to estimate a policy update. RLOO applies a leave-one-out baseline to REINFORCE-style learning from human feedback (Williams, 1992; Ahmadian et al., 2024), while GRPO and DAPO use group-relative learning in reasoning tasks with verifiable rewards (Shao et al., 2024; Yu et al., 2025). In the binary-reward setting studied here, an all-failure group provides no contrast for a raw RLOO update, regardless of its token cost. Increasing the rollout budget offers more opportunities for success but also increases generation cost. This motivates a complementary design problem: organizing a fixed number of on-policy responses into more useful learning signals without changing the expected update.

Structured sampling already offers ways to coordinate language-model generations. Arithmetic Sampling and quasi-random inference encourage diversity through dependent draws (Vilnis et al., 2023; Parashar et al., 2025); QuasiMoTTo extends quasi-Monte Carlo sampling to test-time scaling and examines its implications for RLOO (Li et al., 2026). For a leave-one-out update, however, preserving each response's marginal distribution does not suffice: rewards from other responses enter its baseline, so dependence can introduce nonzero reward–score cross terms. Dependent-gradient estimators address such interactions through appropriate corrections (Dong et al., 2021; Dimitriev & Zhou, 2021). We instead constrain the sampler to preserve the independent law of every trajectory pair, allowing coordination without a coupling-specific correction to raw RLOO.

This constraint leaves a useful distinction between pairwise and joint independence. The established pairwise representation of leave-one-out gradients (Dimitriev & Zhou, 2021; Zhou et al., 2026) makes the mean depend only on two-trajectory laws, while the variance can depend on three- and four-trajectory interactions. We introduce HO-RLOO, an autoregressive sampler that changes these higher-order interactions. Its calibrated mixture assigns rollouts to distinct strata among M probability strata in most groups and to one shared stratum with probability $1/M$, retaining independent randomness within strata. The shared branch restores exactly the pairwise collisions suppressed by the distinct branch. Fresh tokenwise randomization then preserves independence of every pair of complete trajectories under ideal continuous randomness.

Our contribution combines an autoregressive pair-law guarantee (Sections 3–4) with a characterization of the higher-order covariance terms. The aligned sparse-reward analysis identifies a lower-

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.

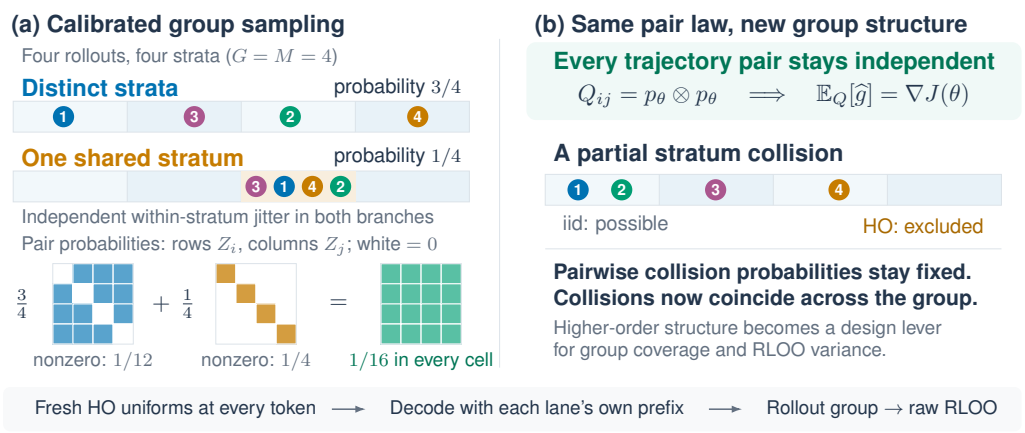


Figure 1: **Same pair laws leave room for coordinated exploration.** (a) For $G = M = 4$, the mixture restores iid pair probabilities; independent jitters give pairwise-independent uniform coordinates. Numbered dots identify rollouts, not rewards. (b) Partial stratum collisions possible under iid are excluded by HO: pairwise collisions instead occur together in the shared branch. Fresh token-wise coupling and prefix-local decoding preserve complete-trajectory pair laws and the raw RLOO mean with fixed logical membership (Corollary 1).

variance regime attaining an established coverage bound (Ramachandra & Natarajan, 2022), and accompanying examples characterize the role of reward geometry (Section 5). A proposal-corrected loss and lagged preconditioner connect the objective to numerical training (Section 5.2). Mathematics and code evaluations measure final policy quality and sampling overhead, while frozen-policy diagnostics examine group reward structure and projected gradient variance (Section 6).

2 Related Work

Dependent sampling and gradient estimation. REINFORCE provides the score-function foundation (Williams, 1992). Beyond independent draws, sampling without replacement yields unbiased discrete-gradient estimators with built-in control variates (Kool et al., 2020), and coupled gradient estimators combine importance sampling with dependent samples (Dong et al., 2021). CARMS expresses leave-one-out updates through pairwise differences and corrects categorical dependence using bivariate probabilities (Dimitriev & Zhou, 2021); the U-statistic view of group-relative gradients exposes their pairwise structure (Zhou et al., 2026). Probability bounds further characterize attainable coverage under pairwise independence (Ramachandra & Natarajan, 2022). We connect these perspectives by enforcing independent trajectory pairs while constructing higher-order dependence. The identity and coverage inequality are established results; our contribution is the autoregressive construction, its cross-kernel variance mechanism, and attainment of the bound in an aligned scalar regime.

Coordinated language-model generation. Arithmetic Sampling couples parallel decoding for diversity (Vilnis et al., 2023), and quasi-random inference organizes multiple language-model samples through structured randomness (Parashar et al., 2025). QuasiMoTTo applies quasi-Monte Carlo sampling to test-time scaling and derives a RLOO pair-law correction under support conditions (Li et al., 2026). HO-RLOO instead makes each complete trajectory pair independent, eliminating a coupling-specific correction; its separate training weight addresses behavior–scoring mismatch (Section 5.2). PAIR reweights inclusion under adaptive rollout allocation (Nomand et al., 2026). Whereas this corrects selection from a candidate group, we design complete-group generation at a fixed rollout count before observing rewards.

Exploration and compute in reinforcement fine-tuning. RLOO, GRPO, and DAPO establish the group-based learning setting (Ahmadian et al., 2024; Shao et al., 2024; Yu et al., 2025). Exploration interventions include lookahead tree-based rollouts (LATR; Xing et al., 2026) and sample-

then-forget optimization (EEPO; Chen et al., 2026). Compute-oriented methods select informative trajectory pairs and dynamically shuffle training data (Shuffle-R1; Zhu et al., 2026), or reuse rollout prefixes (PROS; Huang & Wan, 2026). Polychromic Objectives promotes diversity through vine sampling and modified advantages (Hamid et al., 2026); risk-sensitive RL interpolates between mean and maximum rewards (Jiang et al., 2026). HO-RLOO changes the within-group joint law while preserving the expected raw RLOO gradient. This isolates a sampling-noise intervention from objective modification; Section 6.1 evaluates final policy quality together with sampling overhead.

3 THE PAIRWISE CONSTRAINT

For a prompt $x \sim D$ and response $\tau = (y_1, \dots, y_T)$, $T \le H$, our target is

$$p_\theta(\tau \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t}), \qquad J(\theta) = \mathbb{E}_{x\sim D, \tau\sim p_\theta}[R(x,\tau)]. \tag{1}$$

The reward is fixed given the trajectory; EOS is absorbing. Assume a finite vocabulary and horizon, differentiable positive probabilities on a parameter-independent support, and finite estimator second moments. Below, distributions condition on x and the current policy. Write $s_i = \nabla_\theta \log p_\theta(\tau_i \mid x)$ and $g_x = \mathbb{E}[R_i s_i \mid x]$.

Dependence and baseline bias. For $G \ge 2$ rollouts, raw RLOO (Ahmadian et al., 2024) uses

$$b_{-i} = \frac{1}{G-1}\sum_{j\ne i} R_j, \qquad A_i = R_i - b_{-i}, \qquad \widehat{g} = \frac{1}{G}\sum_i A_i s_i. \tag{2}$$

Preserving each rollout's marginal is insufficient: the baseline introduces cross terms $\mathbb{E}_Q[R_j s_i]$ that need not vanish under a dependent group law Q. The established pairwise representation (Dimitriev & Zhou, 2021; Zhou et al., 2026) exposes a sufficient constraint:

$$\widehat{g} = \frac{1}{K}\sum_{i<j} h(\tau_i, \tau_j), \qquad h(\tau_i,\tau_j) = \frac{1}{2}(R_i - R_j)(s_i - s_j), \quad K = \binom{G}{2}. \tag{3}$$

Matching every trajectory-pair law to iid therefore preserves the mean of every kernel, and hence the expected update (Theorem 3, Appendix A.2). Implementation still uses the linear-time RLOO form.

Admissible group laws. We design within

$$\mathcal{Q}_2(p_\theta) = \{Q : Q_{ij} = p_\theta \otimes p_\theta \text{ for every } i \ne j\}. \tag{4}$$

These laws share the expected raw RLOO gradient while allowing different coverage and covariance. Pairwise independence supplies a reward-independent sufficient constraint that can be enforced before observing outcomes. We focus on raw RLOO to isolate sampling effects at a fixed expected gradient. GRPO's group normalization need not preserve this mean (Appendix A.12); PPO clipping (Schulman et al., 2017) also defines a distinct estimator.

4 CONSTRUCTING PAIRWISE-INDEPENDENT ROLLOUTS

We construct pairwise-independent uniform coordinates, then pass each through its own inverse CDF. The coupling organizes randomness without inspecting rewards or changing token probabilities.

A calibrated stratum mixture. Partition $[0,1)$ into $M \ge G$ equal strata. With probability $1/M$, assign every member one uniformly drawn stratum; otherwise draw G distinct strata uniformly without replacement. This mixing probability restores the stratum collisions that an all-distinct assignment removes from an iid pair. Draw independent jitters $V_i \sim U[0,1)$ in *both* branches and set

$$U_i = \frac{Z_i + V_i}{M}, \qquad \mathbb{P}(\text{shared branch}) = \frac{1}{M}. \tag{5}$$

Independent within-stratum jitters complete the calibrated construction in both branches.

Each pair (U_i, U_j) is independent uniform, although the group is not jointly independent for $G \geq 3$ (Theorem 4, Appendix A.3). We use $M = G$ as the simplest default: it is the smallest admissible stratum count, and the distinct branch is a permutation over all strata. Appendix C.8 studies larger M at fixed G. For $G = 2$, the whole group is iid.

Lifting the construction through autoregressive decoding. At position t, member i has prefix $h_{i,t} = (x, y_{i,<t})$. In a fixed vocabulary order, let $F_{\theta,h}$ be its token CDF. Decode using

$$y_{i,t} = F^{-1}_{\theta,h_{i,t}}(U_{i,t}), \qquad F_{\theta,h}(v) = \sum_{w \leq v} \pi_\theta(w \mid h). \tag{6}$$

Draw a fresh HO group of uniforms at each position, independently of the entire past. Each pair then has the product next-token law, even when its prefixes differ.

Corollary 1 (Complete-trajectory pair law)**.** *With fresh ideal continuous randomness, fixed logical group membership, and deterministic prefix-local decoding, each pair of complete responses has law $p_\theta \otimes p_\theta$. Hence the resulting raw RLOO estimator is unbiased.*

Induction multiplies the conditional product law over positions (Appendix A). Finished members retain their logical random-number positions but can stop model computation, handling variable lengths without outcome-dependent regrouping.

Algorithm 1 One HO rollout group ($M = G$)

Input: Prompt x, policy CDFs $F_{\theta,h}$, group size G, horizon H
Output: Complete or capped responses $\tau_{1:G}$ and behavior log probabilities

Initialize $t \leftarrow 1$, empty responses, and fixed logical lane indices $1, \ldots, G$
while $t \leq H$ and at least one lane is active **do**
  Draw $C_t \sim \text{Bernoulli}(1/G)$ independently of the past
  if $C_t = 1$ **then**
    Draw $Z \sim \text{Uniform}\{0, \ldots, G-1\}$; set $Z_i \leftarrow Z$ for all lanes
  else
    Draw $(Z_1, \ldots, Z_G)$ as a uniform permutation of $\{0, \ldots, G-1\}$
  end if
  Draw independent $V_i \sim U[0, 1)$ for all lanes; set $U_i \leftarrow (Z_i + V_i)/G$
  for all active lanes i **do**
    Append $F^{-1}_{\theta,h_{i,t}}(U_i)$ to τ_i and record its behavior log probability
    Mark lane i finished on EOS; retain its index in subsequent draws
  end for
  $t \leftarrow t + 1$
end while
return Responses $\tau_{1:G}$ and recorded behavior log probabilities

Cost and scope. The extra work per position is a branch, permutation, and G jitters. Coordination operates on probability strata; the decoder maps these coordinates to tokens. The theorem also applies to deterministic represented CDFs, with their behavior laws replacing p_θ. Section 5.2 handles scoring mismatch and finite execution, and Section 6.1 reports training resources and profiles sampler overhead.

5 Higher-order analysis and training

5.1 Variance mechanism and conditional gains

Preserving each pair kernel's law does not preserve dependence *between* kernels. Expanding the covariance of Equation 3 pairs each h_{ij} with every $h_{k\ell}$. Their union contains two, three, or four distinct rollouts. For an exchangeable group, define $a = \text{tr}\,\text{Var}(h_{12})$, $b = \text{tr}\,\text{Cov}(h_{12}, h_{13})$, and $c = \text{tr}\,\text{Cov}(h_{12}, h_{34})$. Grouping by overlap gives

$$\text{tr}\,\text{Var}(\widehat{g}) = \frac{1}{K}\left[a + 2(G-2)b + \binom{G-2}{2}c\right], \qquad K = \binom{G}{2}. \tag{7}$$

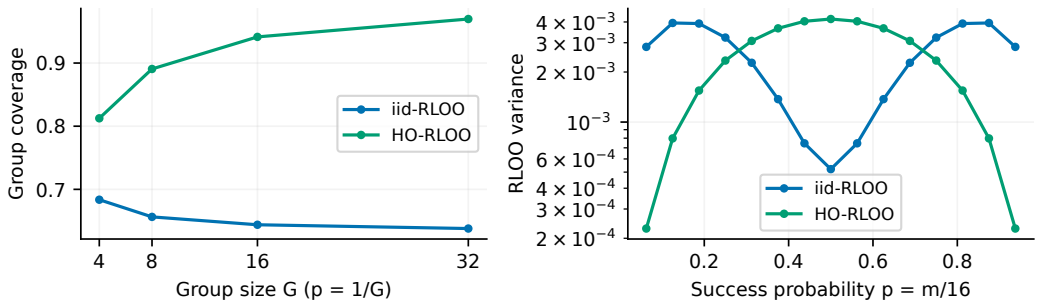


Figure 2: **An exact benefit in the aligned scalar model.** Left: at $p = 1/G$, group coverage increases while the RLOO mean is unchanged. Right: variance across the full grid $p = m/16$; Theorem 2 gives the exact formulas and improvement boundary.

Terms involving nonexistent triples or quadruples are omitted for $G < 3$ or $G < 4$. The pair law fixes a, while the higher-order joint law determines b and c. Under iid sampling $c = 0$; with higher-order dependence, disjoint *kernels* can be correlated. These cross-kernel terms provide the mechanism for changing RLOO variance.

The mechanism is estimator-specific. Fixed-baseline REINFORCE averages one-trajectory terms, whose variance is already fixed by pairwise independence. RLOO instead has an order-two kernel. An order-r kernel needs r-way laws for its mean and up to $2r$-way laws for its second moment (Appendix A.5).

An aligned Bernoulli model. Consider a Bernoulli policy with success probability p, binary reward R, and logit score $s = R - p$. For $N = \sum_i R_i$, Equation 3 reduces to

$$\widehat{g} = \frac{N(G-N)}{G(G-1)}. \tag{8}$$

All-failure and all-success groups give zero contrast; mixed groups give a positive update.

Suppose $p = m/G$, $1 \le m < G$, and the success set occupies m complete strata of the scalar $M = G$ sampler. The distinct branch has $N = m$; the shared branch has $N \in \{0, G\}$. This changes coverage $\mathbb{P}(N > 0)$ and the variance of the quadratic update while preserving $\mathbb{E}N = Gp$ and $\mathrm{Var}(N) = Gp(1-p)$.

Theorem 2 (An aligned sparse-reward benefit)**.** *In the aligned Bernoulli model above, write $v = p(1-p)$. Both iid and HO-RLOO have mean v, and*

$$\mathrm{Var}_{\mathrm{HO}}(\widehat{g}) = \frac{v^2}{G-1}, \qquad \mathrm{Var}_{\mathrm{iid}}(\widehat{g}) = \frac{v}{G}\left[1 - \frac{4G-6}{G-1}v\right]. \tag{9}$$

HO has smaller variance exactly when $v < (G-1)/(5G-6)$. At $p = 1/G$ and $G \ge 4$,

$$C_{\mathrm{HO}} = 1 - \frac{1}{G} + \frac{1}{G^2}, \quad C_{\mathrm{iid}} = 1 - \left(1 - \frac{1}{G}\right)^G, \quad \frac{\mathrm{Var}_{\mathrm{iid}}}{\mathrm{Var}_{\mathrm{HO}}} = G - 4 + \frac{6}{G}. \tag{10}$$

Moreover, C_{HO} attains the maximum coverage among all pairwise-independent Bernoulli groups with marginal success probability $1/G$.

The coverage statement attains a classical pairwise-independence bound (Ramachandra & Natarajan, 2022); the RLOO variance formulas identify when the construction is beneficial. Count distributions, moment calculations, and the attainment proof are in Appendices A.6–A.7.

Benefit regimes. At $G = 16, p = 1/16$, coverage increases from 0.644 to 0.941 and the variance ratio is 12.375. Theorem 2 gives the exact improvement boundary; outside that regime, the variance ordering can reverse (Appendix A.6). The gain depends on reward geometry, with the neutral case characterized in Appendix A.7.

5.2 Training objective and update

For B independently randomized prompt groups, we minimize

$$\mathcal{L}(\theta;\theta_t) = -\frac{1}{BG}\sum_{b=1}^{B}\sum_{i=1}^{G}\mathrm{sg}(w_{b,i}A_{b,i})\log p_\theta(\tau_{b,i}\mid x_b), \tag{11}$$

with raw RLOO advantages. Here sg stops gradients, and $w_i = \mathsf{P}_t(\tau_i \mid x)/\mathsf{Q}_{t,i}(\tau_i \mid x)$ aligns the cached behavior law with the normalized scoring target through a per-trajectory proposal correction. When behavior and target coincide, $w_i = 1$. The implementation retains unclipped, unnormalized weights and a common-support floor. Appendices A.8 and A.11 give the assumptions, proof, and numerical residuals.

The policy stays frozen through sampling and gradient accumulation, followed by one update. Trajectory log probabilities sum over generated tokens, including sampled EOS, without length normalization. All arms use lagged RMS scaling (Equation 42). Theorem 7 connects lower preconditioned variance to an improved one-step smoothness bound; Section 6 evaluates end-to-end learning.

6 Experiments

We evaluate benchmark performance during training and final reasoning quality under a common training-token cap, measure sampling overhead (Section 6.1), examine frozen-policy reward and gradient structure (Section 6.2), and vary group size and estimator design (Appendix C.9).

Models, data, and matched training. We train Qwen3-1.7B and Qwen3-4B (Yang et al., 2025) with rank-32 LoRA adapters (Hu et al., 2022). Every update uses eight mathematics prompts from an 8,192-problem DeepScaleR split (Luo et al., 2025b) and eight code prompts from an 8,060-problem DeepCoder split (Luo et al., 2025a), with 16 responses per prompt. Each domain has 512 disjoint validation problems. A shared non-thinking template, binary verifier, proposal correction, and lagged RMS optimizer are fixed across iid-RLOO, HO-RLOO, and the stratified QuasiMoTTo control QMT-S (Li et al., 2026). Each method uses seeds 7/11/23 at 1.7B and 11/23/47 at 4B, giving 18 training runs. Within each seed cohort, initialization, ordered prompt stream, schedule, and evaluation jobs are matched; the sampler is the intervention. Every run has an eight-million-completion-token budget, with its last complete update at or below that cap selected. Appendix C gives the full configuration.

Structured-sampling control. QMT-S instantiates QuasiMoTTo-style sequence stratification under the common raw-RLOO estimator, including the same trajectory proposal weight, to isolate the sampling intervention. Appendix C.3 specifies this control and derives the pair-support obstruction for direct dependence correction.

Evaluation and uncertainty. All final evaluations use iid decoding. We report pass@1 on MATH-500 (Hendrycks et al., 2021; Lightman et al., 2023), AIME 2024, AIME 2025, and 279 LiveCodeBench-v5 problems dated August 2024–January 2025 (Jain et al., 2024), with 4/32/32/10 samples per problem. Macro weights the mathematics mean and code equally. Main tables report mean $\pm$ sample SD over three training seeds; method contrasts pair the same seed. We report Student-t intervals across seed differences separately from problem-bootstrap intervals conditional on the checkpoints. Final-test accuracy does not select checkpoints.

6.1 Reasoning quality and sampling overhead

Quality under a common token cap. HO-RLOO has the highest mean pass@1 on all four benchmarks at both scales (Table 1). Relative to iid-RLOO, the mathematics mean improves by 2.65 points at 1.7B and 3.02 points at 4B; code improves by 2.37 and 2.53 points. Mean macro gains are 2.51 points (95% paired-seed interval [0.29,4.73]) and 2.78 points ([1.64,3.92]). With iid test decoding in every arm, these contrasts measure the learned policy.

Table 1: **Final reasoning quality at a nominal eight-million-token budget.** Trained rows give mean ± sample SD over three matched seeds: 7/11/23 for 1.7B and 11/23/47 for 4B. All test decoding is iid. The initial-policy row is a shared reference; Macro weights mathematics and LCB equally. QMT-S uses raw RLOO without pairwise correction (Appendix C.3).

Model	Sampler	MATH500	AIME24	AIME25	LCB	Macro
1.7B	Initial policy	72.30	9.48	6.04	13.55	21.41
1.7B	iid-RLOO	76.72±0.68	12.99±0.74	8.65±0.31	16.67±0.57	24.72±0.47
1.7B	QMT-S/RLOO	76.87±0.28	13.85±0.48	9.34±0.22	17.32±0.27	25.34±0.23
1.7B	HO-RLOO	79.05±0.46	16.49±0.69	10.76±0.26	19.03±0.43	27.23±0.43
4B	Initial policy	81.40	19.38	13.12	24.41	31.19
4B	iid-RLOO	84.10±0.58	23.30±0.63	16.49±0.37	27.95±0.40	34.62±0.06
4B	QMT-S/RLOO	84.52±0.55	24.13±0.57	17.19±0.68	28.63±0.22	35.29±0.19
4B	HO-RLOO	86.08±0.49	27.36±0.47	19.51±0.69	30.48±0.52	37.40±0.52

QMT-S improves the mean macro score over iid at both scales, while HO retains a further advantage. The QMT-S contrast characterizes stratified sampling under the common raw RLOO rule (Appendix C.3); HO additionally preserves its expected target under the theorem's assumptions. All three matched seed pairs show a positive HO–iid macro difference, with varying magnitudes. Table A9 reports every run; Table A8 separates seed variability from problem sampling uncertainty.

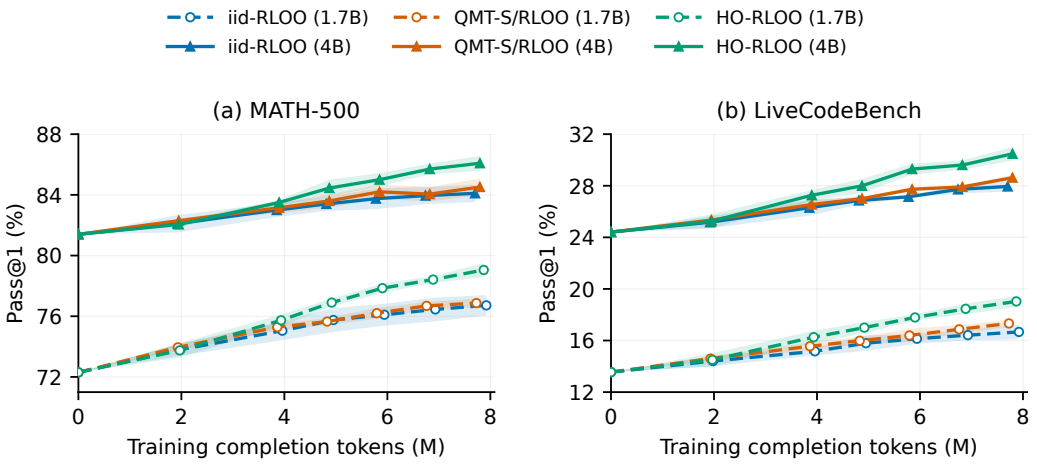


Figure 3: **Benchmark performance during training.** Each panel compares three samplers at 1.7B (dashed) and 4B (solid). Points give the three-seed mean at each checkpoint stage; horizontal coordinates are mean actual training completion tokens. Shading denotes ±1 sample SD across seeds. All evaluations use iid decoding; final points reproduce Table 1.

Benchmark trajectories. Figure 3 evaluates saved checkpoints on MATH-500 and LiveCodeBench using the final-benchmark decoding protocol. The 4B policies start at higher scores, while HO's separation from the controls becomes more pronounced in the second half of training at both scales. With iid decoding at every checkpoint, these trajectories connect the endpoint gains to the evolution of learned policy quality across the training budget.

Execution resources and sampling overhead. All runs use the same eight-million-token cap and select the last complete update within it. Actual endpoint token counts and update counts therefore vary; Table A10 reports these quantities, GPU-hours, throughput, and peak memory for every seed. At 1.7B, mean throughput is 126.16 tokens/s for iid and 126.13 for HO. In a separate profile at the shared initialization, sampler service accounts for 3.18% and 3.31% of forward-plus-sampler time for iid and HO, respectively, compared with 5.38% for QMT-S (Table 2). The small difference in sampler service shows that the coordination can be implemented with little additional overhead in this setting.

Table 2: **Sampler service at the shared 1.7B initialization.** Each arm processes 65,536 action tokens at $G = 16, H = 4096$. Share is sampler time divided by sampler plus forward time.

Sampler	Forward (s)	Sampler (s)	Share (%)
iid-RLOO	307.12	10.09	3.18
QMT-S/RLOO	308.36	17.53	5.38
HO-RLOO	306.85	10.50	3.31

6.2 Connecting the joint distribution to learning signals

Finite random-source checks. The implementation generates FP64 plans with a CPU `torch.Generator` before GPU transfer. An audit of 4,194,304 positions per fixed lane pair gives a maximum stratum-cell discrepancy of 1.21×10^{-4} from $1/256$, consistent with the product-uniform target at the tested resolution. Appendix B.4 reports the complementary conditional-jitter, cross-position, cross-stream, and execution-path checks.

Frozen-policy mechanism. We use the 1.7B seed-7 initialization and halfway iid checkpoint on eight fixed prompts, with 16 groups per prompt and sampler. An independent 64-response iid pool estimates difficulty and the reference gradient. A fixed 64-dimensional projection of the last block's adapter scores measures variance in an unpreconditioned subspace, conditional on this cohort.

Table 3: **Frozen-policy diagnostics,** $G = 16$**.** Each row pools eight prompts, 16 groups per prompt, with equal domain weights. V is the mean within-prompt variance trace of a fixed 64-dimensional projected group gradient; cV averages the prompt-specific token-cost products.

Policy	Sampler	Coverage	Mixed	Reward	V	$cV/10^3$
initial	iid-RLOO	60.16	59.38	18.60	10.83	121.88
initial	QMT-S/RLOO	64.06	63.28	18.51	9.85	111.24
initial	HO-RLOO	65.62	64.06	19.09	9.51	106.79
halfway	iid-RLOO	71.88	71.09	21.44	10.94	125.37
halfway	QMT-S/RLOO	73.44	72.66	21.44	9.96	114.83
halfway	HO-RLOO	73.44	72.66	21.24	9.51	109.91

At initialization, HO raises the mixed-group rate by 4.69 points and lowers the domain-balanced token-cost–variance product by 12.38%. At the halfway policy, the corresponding changes are 1.56 points and 12.33%. The domain and prompt decompositions in Appendix C.7 describe how these aggregate gains arise from changes in group reward structure; the complete histograms retain both uniform-reward extremes.

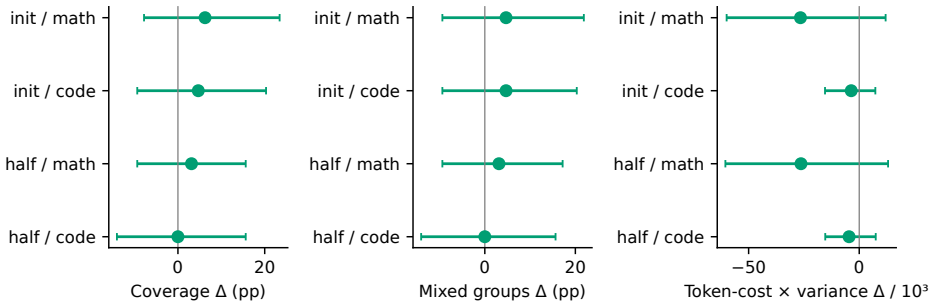


Figure 4: **HO minus iid at frozen policies.** Bars are marginal 95% intervals from 2,000 bootstrap replicates resampling prompts and then complete groups in the seed-7 diagnostic cohort. Reference pools are held fixed. Negative cost–variance differences favor HO.

The diagnostics measure group reward structure and projected variance on the evaluated panel, rather than the full preconditioned variance in Theorem 7. Appendix C.12 states their scope.

6.3 Design choices and rollout diversity

Strata at fixed rollout count. We vary $M/G \in \{1, 2, 4, 8\}$ at fixed $G = 16$, retaining the shared-branch probability $1/M$ and hence the calibrated pair law. Each condition uses the same eight prompts and two frozen policies, with 64 groups per prompt (8,192 responses per condition and policy). Figure A4 in Appendix C.8 separates two design criteria: $M/G = 2$ gives the highest coverage point estimate, while the default $M = G$ gives the lowest projected token-cost–variance among the tested settings, with ratios to iid of 0.926 at initialization and 0.914 halfway. Thus maximizing coverage alone would select a different setting from minimizing projected estimator cost–variance. The appendix reports all measurements and the approach to the iid limit as M increases.

Diversity beyond pairwise collisions. On the original frozen-policy groups, HO increases sequence uniqueness and Distinct-2 while pair-token collision rates remain close to iid (Table 4). All-lane agreement also increases: higher-order coordination need not suppress every form of collision. These descriptive measurements complement the pair-law guarantee by showing how group-level diversity can change while pairwise statistics remain similar. They are not equivalence tests of those statistics.

Table 4: **Rollout diversity and collision structure.** Group averages over eight prompts and 16 groups per prompt at each frozen policy, with $G = 16$. Unique is the fraction of distinct sequences; Distinct-2 is the fraction of distinct bigrams. Token statistics exclude prompt tokens and positions with fewer than two active lanes; all-lane agreement uses positions with all 16 lanes active.

Policy	Sampler	Unique	Distinct-2	Pair collision	All agree
initial	iid-RLOO	0.950	0.679	0.215	0.021
initial	QMT-S/RLOO	0.943	0.681	0.216	0.021
initial	HO-RLOO	0.956	0.704	0.217	0.025
halfway	iid-RLOO	0.943	0.679	0.215	0.021
halfway	QMT-S/RLOO	0.949	0.680	0.216	0.021
halfway	HO-RLOO	0.962	0.705	0.216	0.026

Group-size comparisons at $G \in \{8, 16, 32\}$ and shared-branch and estimator ablations are reported in Appendix C.9. Together with the fixed-G comparison, they distinguish rollout count, higher-order sampling structure, and estimator choice.

7 Conclusion

HO-RLOO preserves the raw RLOO mean while shaping higher-order rollout dependence. The aligned sparse-reward analysis gives exact coverage and variance gains. Under a common training-token cap, domain-balanced pass@1 improves by 2.51 and 2.78 points over iid at two Qwen3 scales, with little additional sampler overhead in the measured profile. These results support higher-order sampling for raw RLOO; broader model families and training scales remain to be evaluated.

REPRODUCIBILITY STATEMENT

Appendix A gives assumptions, complete proofs, and numerical residual bounds. Appendix B specifies the numerical sampler and proposal-corrected training update in pseudocode, together with finite-state and implementation correspondence checks. Appendix C specifies the data, training configuration, evaluation, statistical estimands, and supplementary result tables. Endpoint training resources, independent sampler timing, and diagnostic sampling budgets are reported separately.

AI USE STATEMENT

ChatGPT was used for language editing, LaTeX formatting, reference and consistency checks, assistance with the derivation in Appendix A.12.

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, 2024. URL https://arxiv.org/abs/2402.14740.

Liang Chen, Xueting Han, Qizhou Wang, Bo Han, Jing Bai, Hinrich Schuetze, and Kam-Fai Wong. EEPO: Exploration-enhanced policy optimization via sample-then-forget. In *International Conference on Learning Representations*, pp. 150418–150434, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f39158699fc2f660de36102dd1b61975-Abstract-Conference.html.

Alek Dimitriev and Mingyuan Zhou. CARMS: Categorical-Antithetic-REINFORCE Multi-Sample Gradient Estimator. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13217–13229, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6e16656a6ee1de7232164767ccfa7920-Abstract.html.

Zhe Dong, Andriy Mnih, and George Tucker. Coupled gradient estimators for discrete latent variables. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24498–24508, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cd0b43eac0392accf3624b7372dec36e-Abstract.html.

Jubayer Hamid, Ifdita Orney, Ellen Xu, Chelsea Finn, and Dorsa Sadigh. Polychromic objectives for reinforcement learning. In *International Conference on Learning Representations*, pp. 37356–37381, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3ee63632d4c26ddc3d350c86ffca381c-Abstract-Conference.html.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Baizhou Huang and Xiaojun Wan. PROS: Towards compute-efficient RLVR via rollout prefix reuse. In *International Conference on Learning Representations*, pp. 45546–45562, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4badb55ba9c18bccb9d4146be328a948-Abstract-Conference.html.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code, 2024. URL https://arxiv.org/abs/2403.07974.

Yuhua Jiang, Jiawei Huang, Yufeng Yuan, Xin Mao, YuYue, Qianchuan Zhao, and Lin Yan. Risk-sensitive reinforcement learning for alleviating exploration dilemmas in large language models. In *International Conference on Learning Representations*, pp. 47117–47141, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4e6d14709eae0cbc49a1d19d87fb8b21-Abstract-Conference.html.

Wouter Kool, Herke van Hoof, and Max Welling. Estimating gradients for discrete random variables by sampling without replacement. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rklEj2EFvB.

Michael Y. Li, Anthony Zhan, Kanishk Gandhi, Noah D. Goodman, and Emily B. Fox. Quasi-MoTTo: Quasi-Monte Carlo Test-Time Scaling, 2026. URL https://arxiv.org/abs/2607.01179.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepCoder: A fully open-source 14b coder at O3-mini level, 2025a. URL https://huggingface.co/datasets/agentica-org/DeepCoder-Preview-Dataset. Project release and accompanying preview dataset.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing O1-Preview with a 1.5b model by scaling RL, 2025b. URL https://huggingface.co/datasets/agentica-org/DeepScaleR-Preview-Dataset. Project release and accompanying preview dataset.

Pixel Nomand, Elena Voss, Marcus Hale, and Sofia Reyes. PAIR: Pairwise-Aware Inclusion Reweighting for Adaptive Rollout Allocation in RLVR, 2026. URL https://arxiv.org/abs/2608.11368.

Aditya Parashar, Aditya Vikram Singh, Avinash Amballa, Jinlin Lai, and Benjamin Rozonoyer. Quasi-random Multi-Sample Inference for Large Language Models, 2025. URL https://arxiv.org/abs/2411.06251.

Arjun Ramachandra and Karthik Natarajan. Tight Probability Bounds with Pairwise Independence, 2022. URL https://arxiv.org/abs/2006.00516.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL https://arxiv.org/abs/2402.03300.

Luke Vilnis, Yury Zemlyanskiy, Patrick Murray, Alexandre Tachard Passos, and Sumit Sanghai. Arithmetic Sampling: Parallel Diverse Decoding for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35120–35136. PMLR, 2023. URL https://proceedings.mlr.press/v202/vilnis23a.html.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://link.springer.com/article/10.1007/BF00992696.

Shangyu Xing, Siyuan Wang, Chenyuan Yang, Xin-Yu Dai, and Xiang Ren. Lookahead tree-based rollouts for enhanced trajectory-level exploration in reinforcement learning with verifiable rewards. In *International Conference on Learning Representations*, pp. 31759–31776, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/355d36f5f0922d050d9b6eaa3511d301-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale, 2025. URL https://arxiv.org/abs/2503.14476.

Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Ying Yang, Shijin Gong, and Chengchun Shi. Demystifying Group Relative Policy Optimization: Its Policy Gradient is a U-Statistic, 2026. URL https://arxiv.org/abs/2603.01162.

Linghao Zhu, Yiran Guan, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Bin Qin, Jian Luan, Yuliang Liu, and Xiang Bai. Shuffle-R1: Efficient RL framework for multimodal large language models via data-centric dynamic shuffle. In *International Conference on Learning Representations*, pp. 121372–121393, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c520d1cb9108a27c27227ba039229246-Abstract-Conference.html.

A COMPLETE THEORETICAL DEVELOPMENT

The proofs follow the same constraint-to-construction argument as the main text. We first identify the target gradient and the pairwise terms that RLOO must preserve. We then prove that the stratum mixture satisfies this constraint, first for uniforms and then for complete autoregressive responses. The covariance expansion and Bernoulli example establish what remains free to change. Finally, we transfer the argument to a distinct behavior policy and separate the exact theorem from numerical approximation errors. Unless an expectation explicitly averages over prompts, a prompt x is fixed throughout.

A.1 PROBABILITY SPACE, REGULARITY, AND THE OBJECTIVE

Let $\mathcal{F}_t$ contain the training data specification, the policy and optimizer state before update t, and all randomness used in previous updates. For a fixed prompt x, complete padded trajectories form a finite set when the vocabulary and horizon are finite. EOS is absorbing and adds no score after its first occurrence. Each nonabsorbing conditional distribution is differentiable in θ and positive on a support independent of θ. These conditions justify differentiating finite sums. For a countable or unbounded-horizon extension, the same calculations require an integrable dominating function for differentiation and the displayed second moments.

The terminal reward $R(x, \tau)$ is fixed as a function of the realized trajectory. A deterministic length cap defines the task: the reward is the verifier's output on the capped response. In an environment with tool responses, the trajectory probability includes environment transition kernels independent of θ; the score includes only policy-controlled tokens. Random environment outcomes must be independently generated for each pair conditional on their histories if complete-trajectory pairwise independence is required. The experiments use terminal verifiers without interactive tool transitions.

The trajectory score sums over policy-controlled actions, including sampled EOS:

$$s_\theta(\tau) = \nabla_\theta \log p_\theta(\tau \mid x) = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}). \tag{12}$$

A length-normalized average would define a different estimator. For the conditional objective $J_x(\theta) = \sum_\tau p_\theta(\tau \mid x) R(x, \tau)$,

$$\nabla J_x(\theta) = \sum_\tau R(x, \tau) \nabla p_\theta(\tau \mid x) \tag{13}$$

$$= \sum_\tau p_\theta(\tau \mid x) R(x, \tau) \nabla \log p_\theta(\tau \mid x) = \mathbb{E}[R s_\theta \mid x]. \tag{14}$$

Moreover $\mathbb{E}[s_\theta \mid x] = \sum_\tau \nabla p_\theta(\tau \mid x) = \nabla 1 = 0$. Averaging over the fixed prompt distribution yields Equation 1. For the balanced experiment, $D = \frac{1}{2} D_m + \frac{1}{2} D_c$; eight independent draws from each domain provide an unbiased stratified estimate of this same objective.

A.2 RLOO ALGEBRA AND MEAN

Theorem 3 (Pairwise independence suffices)**.** *If every pair of distinct responses has law $p_\theta(\cdot \mid x) \otimes p_\theta(\cdot \mid x)$, then $\mathbb{E}_Q[\widehat{g} \mid x] = g_x$, irrespective of the higher-order group law.*

For any group law with marginals p_θ, expanding the baseline gives

$$\mathbb{E}_Q \widehat{g} = g_x - \frac{1}{G(G-1)} \sum_{i \neq j} \mathbb{E}_Q[R_j s_i]. \tag{15}$$

Thus correct marginals alone do not control the baseline cross terms.

The first step is algebraic and requires no independence assumption. Write each leave-one-out advantage as an average of reward differences:

$$A_i = R_i - \frac{1}{G-1} \sum_{j \neq i} R_j = \frac{1}{G-1} \sum_{j \neq i} (R_i - R_j). \tag{16}$$

Substituting into the group estimator gives a sum over *ordered* pairs. For each unordered pair $\{i, j\}$, combine the two terms with opposite index orders:

$$\begin{aligned}\widehat{g} &= \frac{1}{G(G-1)} \sum_i \sum_{j \neq i} (R_i - R_j) s_i \\ &= \frac{1}{G(G-1)} \sum_{i<j} \left[(R_i - R_j) s_i + (R_j - R_i) s_j \right] \\ &= \frac{1}{G(G-1)} \sum_{i<j} (R_i - R_j)(s_i - s_j). \end{aligned} \tag{17}$$

Thus the score difference $s_i - s_j$ comes from combining the two directions of a comparison; it is not an additional approximation. Since $2\binom{G}{2} = G(G-1)$, defining $h_{ij} = \frac{1}{2}(R_i - R_j)(s_i - s_j)$ gives Equation 3, with no unspecified constant. As an independent check, if $S_R = \sum_i R_i$ and $S_s = \sum_i s_i$, both numerators equal

$$\sum_{i<j} (R_i - R_j)(s_i - s_j) = (G-1) \sum_i R_i s_i - \sum_{i \neq j} R_i s_j = G \sum_i R_i s_i - S_R S_s. \tag{18}$$

The second step uses the sampling law. Under the hypotheses of Theorem 3,

$$\mathbb{E}[R_i s_i] = g_x, \qquad \mathbb{E}[R_j s_i] = \mathbb{E}[R_j]\mathbb{E}[s_i] = 0 \quad (i \neq j). \tag{19}$$

Consequently $\mathbb{E}h_{ij} = g_x$ and averaging the $\binom{G}{2}$ kernels preserves the target gradient. This proves the theorem without an assumption on triples or larger subsets. The constraint fixes the *distribution* of each pair kernel, not its value on every realized group. Nor does pairwise independence make s_i independent of a nonlinear statistic of all other rewards: the proof works because the leave-one-out baseline is a linear sum of cross terms.

A.3 Uniform construction for arbitrary integers

Theorem 4 (Uniform pair law)**.** *Every U_i in Equation 5 is uniform on $[0, 1)$ and every pair (U_i, U_j) has the independent uniform law. For $G \geq 3$, the complete group is not jointly independent.*

The shared-branch probability follows from the pair constraint. If this branch has probability λ, the probability that a specified pair occupies the same specified bin a is λ/M. The iid target assigns that event probability $1/M^2$, so necessarily $\lambda = 1/M$. We now check that this choice also gives the correct probabilities for unequal bins.

For $M \geq G$, in the shared branch each ordered pair of bins (a, a) has unconditional probability M^{-2}. In the distinct branch, any specified unequal ordered pair occurs with conditional probability $1/[M(M-1)]$: the remaining $G-2$ distinct entries cancel from the numerator and denominator of the uniform ordered-without-replacement distribution. The unconditional probability is again M^{-2}. Therefore

$$\mathbb{P}(Z_i = a, Z_j = b) = M^{-2} \qquad (a, b \in \{0, \ldots, M-1\}). \tag{20}$$

Conditionally on their bins, U_i and U_j have independent densities M on their respective intervals, because V_i, V_j are independent even in the shared branch. The resulting joint density is $M^{-2}M^2 = 1$ on every pair of strata. Boundaries have zero probability. This proves both uniform marginals and pairwise independence. For $G \geq 3$, a partial collision of bins is impossible in the construction but possible under iid uniforms, establishing a nonproduct group law. For $G = 2$, the pair is the entire group and the construction is iid.

The branch probability is therefore a consequence of pairwise independence, not an outcome-tuned repulsion hyperparameter. The main choice $M = G$ makes the distinct branch a full random permutation and requires no finite-field or prime-size restriction. Independence within a shared bin is equally necessary: setting $V_i = V_j$ there would place positive mass on $U_i = U_j$, which independent continuous uniforms cannot have.

A.4 INDUCTION FOR AUTOREGRESSIVE TRAJECTORIES

The uniform-pair theorem must hold conditionally at each step, not just after pooling token positions. This is why the construction draws fresh randomness at every position. Fix any $i \neq j$. At time u, conditional on the entire group past, the freshly generated pair $(U_{i,u}, U_{j,u})$ is independent uniform, with a law that does not depend on that past. The token maps are functions of each member's own prefix. Thus

$$\mathbb{P}(y_{i,u} = a, y_{j,u} = b \mid h_{1:G,u}) \tag{21}$$

$$= \pi_\theta(a \mid h_{i,u})\pi_\theta(b \mid h_{j,u}). \tag{22}$$

Marginalizing irrelevant histories leaves the same expression conditional on the two prefixes. At $u = 1$, it gives the desired product for the first pair of tokens. If the two prefixes have the product autoregressive law through $u - 1$, multiplying by Equation 22 gives that law through u. Induction to H proves Corollary 1.

The argument also covers a terminated member by assigning its absorbing transition probability one. Retaining fixed logical lane indices ensures that active members continue to receive the coordinates of the specified G-member construction. A scheduler may omit model computation for terminated lanes. It must not replace the randomization rule with an unrecorded, outcome-dependent group construction. Fixed token-ID ordering, dropout disabled, and unchanged model weights during the batch make the decoder a deterministic prefix-local map once the uniforms are supplied.

A.5 WHY THIRD- AND FOURTH-ORDER MOMENTS MATTER

The mean examines one kernel at a time; the variance also examines products of two kernels. With $K = \binom{G}{2}$,

$$\operatorname{tr} \operatorname{Var}(\widehat{g}) = \frac{1}{K^2} \sum_{i<j} \sum_{k<\ell} \operatorname{tr} \operatorname{Cov}(h_{ij}, h_{k\ell}). \tag{23}$$

For vector-valued random variables, $\operatorname{tr} \operatorname{Cov}(X, Y) = \mathbb{E}\langle X - \mathbb{E}X, Y - \mathbb{E}Y\rangle$. There are three possible overlap patterns: identical pairs inspect two rollouts; pairs with one common member inspect three; disjoint pairs inspect four. Under exchangeability their covariance traces are respectively a, b (when $G \geq 3$), and c (when $G \geq 4$).

For a fixed pair $\{i, j\}$, there is one identical pair, $2(G - 2)$ other pairs sharing an endpoint, and $\binom{G-2}{2}$ disjoint pairs. The double sum counts ordered choices of the first and second pair, so these counts already include both covariance orders: no additional factor of two is needed. Multiplying by the K choices of the first pair and dividing by K^2 proves Equation 7. Terms requiring more than G distinct members are absent. The pair constraint fixes a, but does not fix the two types of cross-kernel covariance.

Proposition 5 (Estimator-order principle)**.** *Let $\widehat{u}$ be an average of an integrable kernel k over size-r subsets of a group. If every r distinct group members have the iid target law, then $\mathbb{E}\widehat{u}$ equals the iid expectation. If the kernel is square-integrable, its second moment is determined by joint laws of at most $2r$ distinct members. In particular, matching all those joint laws matches the iid covariance.*

Proof. Linearity reduces the mean to the law of one r-member kernel. Expanding $\widehat{u}\widehat{u}^\top$ gives products of two kernels; their union contains at most $2r$ members. Each expectation is therefore determined by the law on that union. Summing proves the statement, including overlapping subsets. □

For a fixed-baseline estimator $G^{-1}\sum_i (R_i - b)s_i$, each summand depends on one trajectory. Pairwise independence makes all cross covariances zero and leaves the individual covariance unchanged. Its variance is consequently the same as under iid sampling. The proposed variance mechanism specifically concerns estimators such as RLOO whose sample-dependent baseline produces a kernel of higher order.

A.6 EXACT BERNOULLI CALCULATIONS

Let $R \sim \text{Bernoulli}(p)$ and parameterize p by its logit θ, so $p = (1 + e^{-\theta})^{-1}$. Then $J(\theta) = p$, $\nabla J = p(1 - p)$, and differentiating $R \log p + (1 - R)\log(1 - p)$ gives $s = R - p$. Hence

$s_i - s_j = R_i - R_j$. For binary rewards $(R_i - R_j)^2$ is one exactly when the pair disagrees. If $N = \sum_i R_i$, there are $N(G-N)$ unordered disagreeing pairs, so

$$\widehat{g} = \frac{N(G-N)}{G(G-1)}. \tag{24}$$

Now impose the alignment assumption: the success set occupies m complete strata, with $p = m/G$ and $1 \le m < G$. In the distinct branch, one draw lands in every stratum, so $N = m$ deterministically. In the shared branch all draws have the same binary reward, although their within-bin jitters remain independent. Thus

$$\mathbb{P}_{\mathrm{HO}}(N = n) = \begin{cases} 1 - 1/G, & n = m, \\ p/G, & n = G, \\ (1-p)/G, & n = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Writing $v = p(1-p)$, Equation 24 is $Gv/(G-1)$ in the distinct branch and zero in the shared branch. The estimator therefore has only two possible values, making its first two moments explicit:

$$\mathbb{E}_{\mathrm{HO}}\widehat{g} = \frac{G-1}{G}\frac{m(G-m)}{G(G-1)} = p(1-p) = v, \tag{26}$$

$$\mathbb{E}_{\mathrm{HO}}\widehat{g}^2 = \frac{G-1}{G}\left(\frac{Gv}{G-1}\right)^2 = \frac{Gv^2}{G-1}, \tag{27}$$

which gives $\mathrm{Var}_{\mathrm{HO}}\,\widehat{g} = v^2/(G-1)$.

For iid rewards $N \sim \mathrm{Binomial}(G, p)$. Define the falling factorial $(N)_k = N(N-1)\cdots(N-k+1)$. It counts ordered selections of k distinct successes, so $\mathbb{E}(N)_k = (G)_k p^k$. Expanding the fourth-degree numerator in this basis gives

$$N^2(G-N)^2 = (N)_4 + (6-2G)(N)_3 + (G^2 - 6G + 7)(N)_2 + (G-1)^2 N. \tag{28}$$

Taking expectations term by term and dividing by $G^2(G-1)^2$ gives

$$\mathbb{E}_{\mathrm{iid}}\widehat{g}^2 = v^2 + \frac{v}{G} - \frac{4G-6}{G(G-1)}v^2. \tag{29}$$

Subtracting v^2 yields Equation 9. Subtracting the HO variance from the iid variance gives

$$\mathrm{Var}_{\mathrm{iid}} - \mathrm{Var}_{\mathrm{HO}} = \frac{v}{G}\left[1 - \frac{5G-6}{G-1}v\right]. \tag{30}$$

For $0 < p < 1$ this is positive precisely under the criterion in Theorem 2. At $p = 1/G$, $v = (G-1)/G^2$, so

$$\mathrm{Var}_{\mathrm{HO}} = \frac{G-1}{G^4}, \qquad \mathrm{Var}_{\mathrm{iid}} = \frac{(G-1)(G^2-4G+6)}{G^5}. \tag{31}$$

Their ratio is $G - 4 + 6/G$. The strict gain starts at $G = 4$; at $G = 3$ the variances agree. At $p = 1/2$ for even G, the HO-to-iid variance ratio is $G/2$. These values locate the operating region of the construction rather than a uniform variance ordering.

A.7 Coverage and the role of alignment

Coverage is $\mathbb{P}(N > 0)$, not the expected success count $\mathbb{E}N$. Pairwise independence fixes $\mathbb{E}N$ and $\mathbb{E}[N(N-1)]$, but these two moments need not determine the probability of an all-failure group. When $1 \le m < G$, the distinct branch always contains a success and the shared branch succeeds with probability p. Thus $C_{\mathrm{HO}} = 1 - 1/G + p/G$. At $p = 1/G$ this gives Equation 10. Under iid sampling $C_{\mathrm{iid}} = 1 - (1-p)^G$.

For any pairwise-independent Bernoulli group with common success probability p, the classical coverage bound (Ramachandra & Natarajan, 2022) has the useful form

$$\mathbb{P}(N > 0) \le \min\{1, Gp - (G-1)p^2\}. \tag{32}$$

Indeed, for $N \in \{0, \dots, G\}$, $\mathbf{1}\{N > 0\} \le N - N(N-1)/G$. The first two factorial moments are Gp and $G(G-1)p^2$ under pairwise independence. Taking expectations proves the bound. At $p = 1/G$, the HO construction attains it, because its success counts lie in $\{0, 1, G\}$, the equality cases of the pointwise inequality.

More generally, let a_j be the conditional success probability within scalar stratum j, with $p = G^{-1}\sum_j a_j$. Independent jitter gives

$$C_{\mathrm{HO}} = \left(1 - \frac{1}{G}\right)\left[1 - \prod_{j=1}^{G}(1 - a_j)\right] + \frac{1}{G}\left[1 - \frac{1}{G}\sum_{j=1}^{G}(1 - a_j)^G\right]. \tag{33}$$

If every $a_j = p$, both branches induce the iid Bernoulli reward law and coverage equals iid coverage. Thus a small marginal success probability alone does not establish a gain. Equation 33 describes a scalar or ideal sequence-coordinate construction; it is not automatically the coverage formula for fresh tokenwise randomization. The latter's reward geometry is measured in Section 6.2.

A.8 Proposal-corrected loss and target preservation

At update t, let $\mathsf{P}_t = \mathsf{P}_{\theta_t}$ denote the target policy in a normalized, differentiable scoring model and define $g_t = \nabla_\theta \mathbb{E}_{x\sim D, \tau\sim \mathsf{P}_\theta}[R(x,\tau)]|_{\theta_t}$. The implementation uses a canonical batch-one, fixed-H, no-cache scorer. Lane i's cached behavior law $\mathsf{Q}_{t,i}$ is defined by represented CDF cell widths under ideal continuous coordinates. Prefix-local decoding gives behavior pair laws $\mathsf{Q}_{t,i} \otimes \mathsf{Q}_{t,j}$ by Corollary 1.

For prompt group b, the trajectory correction is

$$\lambda_{b,i} = \sum_{u \le T_{b,i}} \left[\log p_t(y_{b,i,u} \mid h_{b,i,u}) - \log q_{t,i}(y_{b,i,u} \mid h_{b,i,u})\right], \qquad w_{b,i} = e^{\lambda_{b,i}}. \tag{34}$$

Both terms use the same generated-token mask, including sampled EOS. The weights are detached, neither clipped nor group-normalized.

Theorem 6 (Target preservation under proposal correction)**.** *Assume the normalized differentiable target model above, $\mathsf{P}_t \ll \mathsf{Q}_{t,i}$, and behavior pair laws $\mathsf{Q}_{t,i} \otimes \mathsf{Q}_{t,j}$. With exact trajectory ratios and scores, the negative gradient of Equation 11 at θ_t satisfies $\mathbb{E}_t \bar{g}_t = g_t$, where $\mathbb{E}_t$ conditions on the pre-batch history $\mathcal{F}_t$.*

The preceding results sample from the policy whose score appears in the estimator. We now allow the behavior law and that target law to differ, while retaining pairwise-independent behavior. Fix a prompt and update, suppress t, and let P belong to the normalized differentiable fixed-shape scoring family. Under the ideal continuous random-source model, logical lane i is sampled from the cached represented-CDF behavior law Q_i. Define

$$s_i = \nabla_\theta \log p_\theta(\tau_i \mid x)|_{\theta_t}, \qquad w_i = \frac{d\mathsf{P}}{d\mathsf{Q}_i}(\tau_i), \qquad A_i = R_i - \frac{1}{G-1}\sum_{j\neq i} R_j. \tag{35}$$

The derivative of Equation 11, with w_i and A_i stopped, is the proposal-corrected group estimator

$$\widehat{g}_w = \frac{1}{G}\sum_{i=1}^{G} w_i A_i s_i. \tag{36}$$

The Radon–Nikodym notation includes discrete trajectories: then $w_i(\tau) = \mathsf{P}(\tau)/\mathsf{Q}_i(\tau)$ on the target support. The fixed horizon makes each ratio the exponential of the masked token-log-probability difference in Equation 34.

For completeness, let Q_h^0 be the categorical law obtained from the monotone, capped represented CDF at history h, let $\mathsf{U}_\mathcal{V}$ be uniform on the vocabulary, and define the implemented continuous-coordinate behavior

$$\mathsf{Q}_h^\alpha = (1-\alpha)\mathsf{Q}_h^0 + \alpha \mathsf{U}_\mathcal{V}, \qquad \alpha = 2^{-28}. \tag{37}$$

Every token then has mass at least $\alpha/|\mathcal{V}|$, hence any finite-logit scoring law is absolutely continuous with respect to Q_h^α. Moreover,

$$\mathrm{TV}(\mathsf{Q}_h^\alpha, \mathsf{Q}_h^0) \le \alpha, \qquad \mathrm{TV}(\mathsf{Q}_{1:T}^\alpha, \mathsf{Q}_{1:T}^0) \le 1 - (1-\alpha)^T \le T\alpha \le H\alpha. \tag{38}$$

The trajectory bound follows by coupling the two autoregressive decoders until the first Bernoulli-α uniform-component draw. At $H = 4096$, it is at most $2^{-16} = 1.525879 \times 10^{-5}$. Proposal correction uses the exact Q^{α} cell widths, so Equation 38 measures the change from the unfloored represented sampler rather than a gradient-bias term.

Proof of Theorem 6. Condition on $\mathcal{F}_t$ and x. Absolute continuity and change of measure give

$$\mathbb{E}_{\mathsf{Q}_i}[w_i R_i s_i] = \mathbb{E}_{\mathsf{P}}[Rs] = g_x, \qquad \mathbb{E}_{\mathsf{Q}_i}[w_i s_i] = \mathbb{E}_{\mathsf{P}}[s] = 0. \tag{39}$$

For $i \neq j$, the assumed pair product law therefore yields

$$\mathbb{E}_{\mathsf{Q}_i \otimes \mathsf{Q}_j}[w_i R_j s_i] = \mathbb{E}_{\mathsf{Q}_i}[w_i s_i]\mathbb{E}_{\mathsf{Q}_j}[R_j] = 0. \tag{40}$$

Expanding Equation 36, every direct term contributes g_x and every cross term vanishes, so $\mathbb{E}[\widehat{g}_w \mid x, \mathcal{F}_t] = g_x$. Averaging over the fixed prompt distribution (or its unbiased stratified batch) gives $\mathbb{E}_t \overline{g}_t = g_t$ as defined above. No joint law of three or more lanes enters the calculation. □

The same estimator has an exact weighted pair-kernel representation:

$$\begin{aligned} \widehat{g}_w &= \frac{1}{G(G-1)} \sum_{i<j} (R_i - R_j)(w_i s_i - w_j s_j) \\ &= \binom{G}{2}^{-1} \sum_{i<j} \frac{1}{2} (R_i - R_j)(w_i s_i - w_j s_j). \end{aligned} \tag{41}$$

To verify the coefficient, expand the first numerator: each $R_i w_i s_i$ appears $G - 1$ times and the remaining terms are exactly $-\sum_{i \neq j} R_j w_i s_i$, which is $G(G-1)\widehat{g}_w$. Equivalently, repeat the ordered-to-unordered calculation above with s_i replaced by $w_i s_i$. Proposal correction therefore retains an order-two kernel. If behavior and target coincide, $w_i = 1$ and Equations 36–41 reduce to raw RLOO and Equation 3. This includes the idealized policy-CDF sampler without the support floor. Ideal arithmetic alone does not imply equality when the floor is retained: even if the token laws satisfy $\mathsf{Q}_h^0 = \mathsf{P}_h$, generally $(1-\alpha)\mathsf{P}_h + \alpha \mathsf{U}_{\mathcal{V}} \neq \mathsf{P}_h$.

The order-two argument does not require identical lane laws. Its covariance is always the double sum over weighted kernels. The reduction of that sum to the three coefficients a, b, c in Equation 7 additionally requires exchangeability; it should not be imposed on arbitrary lane-specific numerical laws.

This theorem applies to iid and HO because their represented pair laws are products under the ideal continuous random-source premise of Corollary 1. A finite PRNG law is handled by the separate residual below; it is not silently identified with a continuous uniform. QMT-S deliberately coordinates whole sequences and is evaluated as a dependent-sampling control; neither Theorem 6 nor its zero-bias conclusion is assigned to that arm.

A.9 Conditional progress

The implementation uses lagged RMS scaling with state v_{t-1} fixed before the current batch:

$$\begin{aligned} d_t &= \operatorname{clip}\left((\sqrt{v_{t-1}} + \epsilon)^{-1}, d_{\min}, d_{\max}\right), \qquad P_t = \operatorname{diag}(d_t), \\ \theta_{t+1} &= \theta_t + \eta P_t \overline{g}_t, \qquad v_t = \rho v_{t-1} + (1-\rho)\overline{g}_t^{\odot 2}. \end{aligned} \tag{42}$$

Here ρ is the second-moment decay, $\epsilon > 0$ stabilizes the reciprocal, $\odot 2$ denotes componentwise squaring, and $0 < d_{\min} \leq d_{\max}$ bound each scale. The order matters: the parameter update uses the old moment, not the current batch's squared gradient. The argument below applies to any positive diagonal, history-measurable preconditioner with these bounds. Target preservation connects sampling to optimization through the noise of the actual update, not through coverage alone.

Theorem 7 (Conditional optimization progress)**.** *Suppose J has L-Lipschitz gradient, P_t is $\mathcal{F}_t$-measurable, and $\mathbb{E}_t \overline{g}_t = g_t$. For $V_t = \mathbb{E}_t \|P_t(\overline{g}_t - g_t)\|^2$,*

$$\mathbb{E}_t J(\theta_{t+1}) \geq J(\theta_t) + \eta g_t^{\top} P_t g_t - \frac{L\eta^2}{2}\left(\|P_t g_t\|^2 + V_t\right). \tag{43}$$

For an L-smooth function, the two-sided smoothness inequality gives

$$J(\theta_t + \Delta) \geq J(\theta_t) + g_t^\top \Delta - \frac{L}{2}\|\Delta\|^2. \tag{44}$$

Set $\Delta = \eta P_t \overline{g}_t$ and condition on $\mathcal{F}_t$. Because P_t is measurable with respect to $\mathcal{F}_t$ and $\mathbb{E}_t \overline{g}_t = g_t$, the linear term becomes $\eta g_t^\top P_t g_t$. Writing $\overline{g}_t = g_t + \xi_t$ with $\mathbb{E}_t \xi_t = 0$ gives

$$\mathbb{E}_t\|P_t \overline{g}_t\|^2 = \|P_t g_t\|^2 + \mathbb{E}_t\|P_t \xi_t\|^2. \tag{45}$$

This proves Equation 43. A long-run consequence requires additional bounds: suppose $J \leq J^*$, $V_t \leq \sigma_P^2$, $0 < d_{\min} \leq d_{t,k} \leq d_{\max}$, and $\eta \leq d_{\min}/(L d_{\max}^2)$. Then, after T updates,

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\|g_t\|^2 \leq \frac{2(J^* - J(\theta_0))}{\eta d_{\min} T} + \frac{L\eta\sigma_P^2}{d_{\min}}. \tag{46}$$

To obtain this bound, the eigenvalue inequalities give $g_t^\top P_t g_t \geq d_{\min}\|g_t\|^2$ and $\|P_t g_t\|^2 \leq d_{\max}^2\|g_t\|^2$. Under the stated step-size condition,

$$\mathbb{E}_t[J(\theta_{t+1}) - J(\theta_t)] \geq \frac{\eta d_{\min}}{2}\|g_t\|^2 - \frac{L\eta^2}{2}\sigma_P^2. \tag{47}$$

Taking unconditional expectations, summing from $t = 0$ to $T - 1$, and using $J(\theta_T) \leq J^*$ proves Equation 46. For independent prompt groups, covariance contributions add under the displayed batch normalization; the target-preservation argument is applied conditionally within every iid or HO group.

The smoothness and moment assumptions specify sufficient conditions for the progress guarantee. The theorem identifies the full preconditioned variance term. Our empirical diagnostics measure unpreconditioned score variance in a fixed adapter subspace (Appendix C); final benchmark performance is measured separately.

At a frozen policy and preconditioner, let V be the preconditioned variance of one complete-group estimator and c its mean cost. Averaging approximately $\mathcal{B}/c$ independent groups at budget $\mathcal{B}$ gives variance $cV/\mathcal{B}$. This is a planning approximation, not an optional-stopping identity.

A.10 Optional trajectory-KL objective

Let a fixed reference p_{ref} be positive on the policy support and define

$$J_\beta(\theta) = \mathbb{E}_{p_\theta} R - \beta \mathbb{E}_{p_\theta} \log \frac{p_\theta(\tau \mid x)}{p_{\mathrm{ref}}(\tau \mid x)}. \tag{48}$$

Differentiating both the sampling probability and the log density ratio gives

$$\nabla J_\beta = \mathbb{E}_{p_\theta}\left[\left(R - \beta \log \frac{p_\theta}{p_{\mathrm{ref}}}\right) s_\theta\right] - \beta \mathbb{E}_{p_\theta} s_\theta \tag{49}$$

$$= \mathbb{E}_{p_\theta}[F_\theta s_\theta], \qquad F_\theta = R - \beta \log(p_\theta / p_{\mathrm{ref}}). \tag{50}$$

At θ_t, compute F_i with the target scorer and reference trajectory log probabilities, detach it, and replace R_i by F_i in the advantages. Multiplying each resulting term by w_i then gives the exact gradient of J_β under the same assumptions as Theorem 6. The trajectory log ratio is a sum over the same action-token mask as the score. The θ-dependence of F_θ is already accounted for by the vanishing extra score term and is not differentiated again through the surrogate loss. This is a KL of full trajectory distributions, not an unspecified token-average penalty. The experiments set $\beta = 0$.

A.11 Finite-representation residuals and biased progress

The target-preservation theorem is an exact statement about normalized trajectory laws, their target score, and ideal continuous random coordinates. A finite implementation can depart from these premises in different ways. We separate errors in the density ratio, defects in the behavior pair law, changes induced by a finite random source, and approximation to the target score; their effects then enter a common biased-progress inequality. A small error in a selected token probability alone does not establish these premises.

1. Approximate density ratios with the pair law held fixed. First retain the law and score assumptions of Theorem 6, but replace the exact ratio by $\widetilde{w}_i$. This includes the deliberate omission of correction, $\widetilde{w}_i = 1$.

Proposition 8 (Finite-representation residual)**.** *Assume the hypotheses of Theorem 6, let rewards lie in an interval of width Δ_R, and let P_t be any history-measurable linear preconditioner. If $\widetilde{g} = G^{-1}\sum_i \widetilde{w}_i A_i s_i$, then*

$$\left\|P_t\big(\mathbb{E}_t\widetilde{g} - g_t\big)\right\| \le \frac{\Delta_R}{G}\sum_{i=1}^{G}\mathbb{E}_{\mathsf{Q}_i}\left[|\widetilde{w}_i - w_i|\,\|P_t s_i\|\right]. \tag{51}$$

If $|\log\widetilde{w}_i - \log w_i| \le \epsilon_i$ on the behavior support, the ith expectation is at most $(e^{\epsilon_i}-1)\mathbb{E}_{\mathsf{P}}\|P_t s\|$. In particular, separate numerator and denominator log-density errors bounded by ϵ_i^P and ϵ_i^Q give $\epsilon_i \le \epsilon_i^P + \epsilon_i^Q$.

Proof. Subtract Equation 36 from $\widetilde{g}$. Since both R_i and its leave-one-out average lie in the same reward interval, $|A_i| \le \Delta_R$. The triangle inequality gives Equation 51; Theorem 6 identifies the expectation of the exact-weight estimator with g_t. The logarithmic condition implies $|\widetilde{w}_i - w_i| \le w_i(e^{\epsilon_i}-1)$, and change of measure completes the second claim. □

For the uncorrected executable estimator ($\widetilde{w}_i = 1$), Cauchy–Schwarz gives the explicit corollary

$$\left\|P_t\big(\mathbb{E}_t\widehat{g}_1 - g_t\big)\right\| \le \frac{\Delta_R}{G}\sum_{i=1}^{G}\sqrt{\chi^2(\mathsf{P}\|\mathsf{Q}_i)\,\mathbb{E}_{\mathsf{Q}_i}\|P_t s_i\|^2}, \tag{52}$$

where $\chi^2(\mathsf{P}\|\mathsf{Q}_i) = \mathbb{E}_{\mathsf{Q}_i}(w_i - 1)^2$. The direction of the divergence follows from sampling under Q_i and reweighting to P.

2. A pair-law defect with the marginals held fixed. Pairwise independence is a separate premise from marginal score correspondence. If an actual pair law M_{ij} has exact marginals $\mathsf{Q}_i, \mathsf{Q}_j$ and $\delta_{ij} = \mathrm{TV}(\mathsf{M}_{ij}, \mathsf{Q}_i \otimes \mathsf{Q}_j)$, then the proof of Theorem 6 acquires at most

$$\frac{\Delta_R}{G(G-1)}\sum_{i\ne j} K_i\delta_{ij}, \qquad K_i = \operatorname*{ess\,sup}_{\mathsf{Q}_i} w_i\|P_t s_i\|, \tag{53}$$

in preconditioned bias. To see the constant, center R_j at the midpoint of its range; exact marginals leave the covariance term unchanged, and the standard total-variation expectation bound contributes $2(\Delta_R/2)K_i\delta_{ij}$. Equation 53 is a sensitivity statement, not a claim that the QMT-S pair defect is small.

3. Finite-source changes to both marginals and pairs. The finite random source can also perturb marginals. Let $\mathsf{Q}_i^{\mathrm{rng}}$ and $\mathsf{M}_{ij}^{\mathrm{rng}}$ be the trajectory marginal and pair law induced when the pinned PRNG seed is randomized, while Q_i remains the continuous-uniform represented-CDF law whose cell widths are logged. Assume the relevant supports lie where w_i and s_i are finite, and define

$$\alpha_i = \mathrm{TV}(\mathsf{Q}_i^{\mathrm{rng}}, \mathsf{Q}_i), \qquad \gamma_{ij} = \mathrm{TV}(\mathsf{M}_{ij}^{\mathrm{rng}}, \mathsf{Q}_i^{\mathrm{rng}} \otimes \mathsf{Q}_j^{\mathrm{rng}}), \qquad K_i^\star = \sup_\tau w_i(\tau)\|P_t s_i(\tau)\|. \tag{54}$$

If $\widehat{g}_w^{\mathrm{rng}}$ uses the exact represented-law weights on PRNG samples, then

$$\left\|P_t\big(\mathbb{E}_t\widehat{g}_w^{\mathrm{rng}} - g_t\big)\right\| \le \frac{\Delta_R}{G}\sum_i K_i^\star\alpha_i + \frac{\Delta_R}{G(G-1)}\sum_{i\ne j}K_i^\star(\gamma_{ij}+\alpha_i+\alpha_j). \tag{55}$$

Indeed, RLOO is unchanged after centering rewards at the midpoint of their range. The direct terms then use the marginal TV bound, while the cross terms use $\mathrm{TV}(\mathsf{M}_{ij}^{\mathrm{rng}}, \mathsf{Q}_i\otimes\mathsf{Q}_j) \le \gamma_{ij}+\alpha_i+\alpha_j$ and the same $2\|f\|_\infty\mathrm{TV}$ inequality as Equation 53. Thus the proposal ratio is exact for the represented-CDF model without claiming that a discrete PRNG is bit-exact continuous randomness. If $\widetilde{w}_i$ is used on PRNG samples, the first triangle term is at most $\Delta_R G^{-1}\sum_i \mathbb{E}_{\mathsf{Q}_i^{\mathrm{rng}}}[|\widetilde{w}_i - w_i|\|P_t s_i\|]$; adding it to Equation 55 gives the combined residual. Proposition 8 is the represented-law specialization.

4. Approximation to the target score. The preceding bounds retain the exact score s_i of the normalized differentiable target family. To account separately for a computed score $\widetilde{s}_i = s_i + e_i$, condition on $\mathcal{F}_t$ and let M be the actual joint sampling law of a prompt and its G responses. The same argument applies at a fixed prompt, with its conditional law and target gradient in place of M and g_t. Under this law, use the same applied weights $\widetilde{w}_i$ and raw advantages A_i to define

$$\widetilde{g}_s = \frac{1}{G}\sum_i \widetilde{w}_i A_i \widetilde{s}_i, \qquad \widetilde{g}_0 = \frac{1}{G}\sum_i \widetilde{w}_i A_i s_i. \tag{56}$$

Here e_i is the error relative to the target score, not a change in the definition of the target policy. If the displayed expectations are finite, then

$$\begin{aligned}\left\|P_t\big(\mathbb{E}_{\mathsf{M}}\widetilde{g}_s - g_t\big)\right\| \le \left\|P_t\big(\mathbb{E}_{\mathsf{M}}\widetilde{g}_0 - g_t\big)\right\| \\ + \frac{\Delta_R}{G}\sum_i \mathbb{E}_{\mathsf{M}}\left[|\widetilde{w}_i|\,\|P_t e_i\|\right].\end{aligned} \tag{57}$$

Indeed, $\widetilde{g}_s - \widetilde{g}_0 = G^{-1}\sum_i \widetilde{w}_i A_i e_i$, so the bound follows from $|A_i| \le \Delta_R$ and the triangle inequality. No independence between the weights, rewards, and score errors is required. The first term is bounded by the preceding results when their corresponding law assumptions hold; the second is the additional score-approximation contribution evaluated under the *actual* sampling law. Thus weight approximation and score approximation are distinct errors even when both originate from the same finite computation.

5. Propagating the remaining bias to the optimization step. Let a general implemented estimator have conditional bias $b_t = \mathbb{E}_t\widetilde{g}_t - g_t$, centered noise $\zeta_t = \widetilde{g}_t - \mathbb{E}_t\widetilde{g}_t$, $\beta_t = \|P_t b_t\|$, and $V_t = \mathbb{E}_t\|P_t\zeta_t\|^2$. The same smoothness step gives the exact biased-progress form

$$\mathbb{E}_t[J(\theta_{t+1}) - J(\theta_t)] \ge \eta g_t^\top P_t(g_t + b_t) - \frac{L\eta^2}{2}\left(\|P_t(g_t + b_t)\|^2 + V_t\right) \tag{58}$$

$$\ge \eta d_{\min}\|g_t\|^2 - \eta\beta_t\|g_t\| - \frac{L\eta^2}{2}\left[(d_{\max}\|g_t\| + \beta_t)^2 + V_t\right]. \tag{59}$$

The squared mean therefore contains the bias; omitting it would overstate progress. For exact proposal correction and exact target scores under iid or HO represented behavior and the theorem's random-source premise, $b_t = 0$ and Equation 58 reduces to Theorem 7. Otherwise Equations 51, 53, 55, and 57 supply explicit components of β_t.

The preliminary selected-token smoke checks token identity, masking, and selected-probability discrepancies, but does not by itself establish a trajectory law, support, a pair-product law, or vanishing PRNG-lattice terms. The current run-level audit therefore stores both selected-token log probabilities through EOS/cap, verifies finite common support, reconstructs every trajectory weight, and reports its distribution, effective sample size, grid-mass residual, and corrected–uncorrected gradient difference. Table A1 reports its frozen-batch instance. The pinned generator, seed derivation, and uniform-plan hashes identify the finite-source law; finite-state enumeration can evaluate its analogue of Equation 55. The study in Section 6.2 repeats the projected-gradient diagnostics across policies and prompts. These measurements test the premises and numerical conditioning of the correction without replacing the theorem or the formal effect estimates.

A.12 Scope under group normalization

GRPO uses a group estimator with standardized reward advantages (Shao et al., 2024). To isolate normalization, fix a prompt and consider the on-policy reward component with trajectory scores, before clipping is active. Define $\bar{R} = G^{-1}\sum_i R_i$ and $S^2 = G^{-1}\sum_i (R_i - \bar{R})^2$. With detached advantages and $\epsilon \ge 0$, the normalized estimator satisfies

$$\widehat{g}_{\mathrm{norm}} = \frac{1}{G}\sum_i \frac{R_i - \bar{R}}{S + \epsilon} s_i = \frac{G-1}{G}\frac{\widehat{g}}{S+\epsilon}, \tag{60}$$

where $\widehat{g}$ is raw RLOO and uniform-reward groups contribute zero, including when $\epsilon = 0$. Group centering alone therefore changes RLOO only by a fixed factor. Division by the random group scale

couples every pair kernel to all rewards, so the normalized terms generally depend on more than two trajectories. Pairwise independence still holds for the sampler, but does not fix the expectation in Equation 60.

A counterexample follows from the aligned Bernoulli model in Appendix A.6. Take a single-token policy with $G = 4$, $p = 1/4$, reward $R_i \in \{0, 1\}$, and logit score $s_i = R_i - p$. Let $N = \sum_i R_i$ and $u_n = n(4 - n)/16$. Then $\widehat{g}_{\text{norm}} = u_N/(\sqrt{u_N} + \epsilon)$. The HO law in Equation 25 has probabilities $3/16$, $3/4$, and $1/16$ at $N = 0, 1, 4$, respectively; iid sampling gives $N \sim \text{Binomial}(4, 1/4)$. Direct summation yields

$$\mathbb{E}_{\text{HO}}\widehat{g}_{\text{norm}} - \mathbb{E}_{\text{iid}}\widehat{g}_{\text{norm}} = \frac{27}{512}\left(\frac{1}{\sqrt{3}/4 + \epsilon} - \frac{1}{1/2 + \epsilon}\right) > 0. \tag{61}$$

For $\epsilon = 0$, the two means are $3\sqrt{3}/16$ and $(30\sqrt{3} + 27)/256$. Thus even equality with the iid normalized update, rather than unbiasedness for the reward gradient, fails in general. The single-token example has no length-normalization ambiguity and establishes the obstruction before clipping or rollout reuse. It does not preclude useful higher-order designs for GRPO; it shows why those designs require analysis of both the mean and variance, whereas raw RLOO isolates sampling effects through Theorem 3.

B Training algorithm and numerical correspondence

This appendix specifies the numerical sampler, proposal-corrected update, and execution checks. Algorithm 2 summarizes one training update; Algorithm 1 gives the calibrated rollout construction. The correspondence table identifies the quantities required by the target-preservation argument.

B.1 Quantities that an implementation must preserve

Mathematical object	Implementation counterpart
$\mathsf{P}_t(\tau \mid x)$	Canonical batch-one, fixed-H, no-cache scoring law through EOS/cap.
$\mathsf{Q}_{t,i}(\tau \mid x)$	Product of full-support cached represented-CDF cell widths under ideal continuous coordinates.
$s_\theta(\tau)$	Sum of target-scoring log-probability derivatives on the action mask.
$\mathsf{Q}_{ij} = \mathsf{Q}_i \otimes \mathsf{Q}_j$	Fresh ideal continuous pairwise uniforms, fixed lanes, and prefix-local cached CDFs for iid/HO.
$w_i = d\mathsf{P}/d\mathsf{Q}_i$	FP64 sum of selected $\log p_t - \log q_{t,i}$, then detached exponential.
A_i	Detached raw leave-one-out reward difference.
$\mathbb{E}_t\bar{g}_t = g_t$	Proposal-corrected iid/HO represented law under the theorem premises.
P_t fixed given $\mathcal{F}_t$	Scale from the previous optimizer state; second moment updated afterward.

B.2 TRAINING UPDATE

Algorithm 2 One proposal-corrected HO-RLOO update

Input: Policy θ_t, previous moment v_{t-1}, and prompts $x_{1:B}$
Output: Updated policy θ_{t+1} and moment v_t

Freeze θ_t and disable dropout throughout sampling and gradient accumulation
for $b = 1, \dots, B$ **do**
  Generate G responses with Algorithm 1 and the represented CDF below
  Retain fixed lanes through EOS and store behavior log masses
  Verify responses to obtain $R_{b,1:G}$; score with the unchanged target policy
  Sum log probabilities through EOS/cap, without length normalization
  for $i = 1, \dots, G$ **do**
    $w_{b,i} \leftarrow \mathrm{sg}[\exp(\log \mathsf{P}_t(\tau_{b,i} \mid x_b) - \log \mathsf{Q}_{t,i}(\tau_{b,i} \mid x_b))]$
    $A_{b,i} \leftarrow \mathrm{sg}[R_{b,i} - (G-1)^{-1} \sum_{j \neq i} R_{b,j}]$
  end for
end for
Accumulate $\overline{g}_t \leftarrow (BG)^{-1} \sum_{b,i} w_{b,i} A_{b,i} \nabla_\theta \log p_\theta(\tau_{b,i} \mid x_b)|_{\theta=\theta_t}$
Form P_t from v_{t-1} using Equation 42
$\theta_{t+1} \leftarrow \theta_t + \eta P_t \overline{g}_t$
$v_t \leftarrow \rho v_{t-1} + (1-\rho)\overline{g}_t^{\odot 2}$
return θ_{t+1}, v_t

Represented sampling law. At each token position, the cached decoder computes FP32 softmax rows. Their represented values are accumulated in FP64, followed by a cumulative-maximum envelope, a cap at one, and terminal-endpoint completion. Mixing this base CDF with the uniform categorical CDF at weight $\alpha = 2^{-28}$ gives full support. Inverse-CDF decoding uses the fresh coordinates from Algorithm 1, and the logged behavior probability is the selected interval's actual adjacent-endpoint width. Finished lanes retain their logical indices. The finite-source plan and its execution checks are specified in Appendix B.4.

Scoring and accumulation. The target scorer uses batch-one, fixed-H, no-cache teacher forcing. The action mask includes sampled EOS and excludes prompt tokens and post-EOS padding. Trajectory log-ratios are accumulated in FP64; weights are detached, unclipped, and not group-normalized. Each complete prompt group remains on one of two ranks. With $N_{\mathrm{loc}} = 128$ trajectories per rank, microbatch loss contributions use the fixed denominator N_{loc}; averaging rank gradients yields the global $1/(BG)$ factor. Advantages are formed from complete groups before microbatching, and the policy is updated only after all gradients have been accumulated. Verifier infrastructure failures trigger re-scoring of the stored completion, without resampling.

Optimizer and decoder requirements. The lagged RMS scale is defined by Equation 42: the parameter update uses the previous second moment, which is updated afterward. Model computation, adapter parameters, gradients, and optimizer state use FP32, using SDPA with TF32 and autocast disabled. The model input template and truncation rules are part of the policy and reward definitions. Decoding requires explicit control of categorical coordinates and a prefix-local token map; assigning related seeds to an opaque generation API does not specify the required coupling.

B.3 EXECUTION CORRESPONDENCE ON THE PINNED MODEL

The production HO and iid samplers were first replayed through the exact Qwen3-1.7B snapshot with FP32 softmax and finite FP64 inverse-CDF coordinates. Across 2,048 action tokens per sampler, the maximum selected-token discrepancy between cached behavior logging and a teacher-forced score pass was 1.030×10^{-4} for iid and 8.595×10^{-5} for HO. Both are below 10^{-3}. This check measures token/path alignment and numerical scale; it does not certify a trajectory law or expected gradient. The two-rank trainer also matched a one-update checkpoint resumed to update two against an uninterrupted two-update run across every recorded tensor and scientific state field. The comparison included RNG state, data cursors, and budget counters.

The full-support proposal-correction audit uses one frozen group of 16 trajectories and 6,179 action tokens. Replaying the same coordinate tape reproduced the tokens, masks, EOS state, decode steps,

and rewards exactly. Support, nonfinite, and underflow violations were all zero. Table A1 reports the weight and gradient diagnostics; suffix/padding perturbation changed no scored value and had gradient cosine one. The check leaves the model and optimizer unchanged.

Table A1: **Current-protocol proposal-correction audit.** These are numerical and implementation checks on a frozen batch, not estimates of training effectiveness.

Quantity	Observed value
Normalized effective sample size	0.999999969344
99th percentile of $\lvert \log w \rvert$	$4.45830159 \times 10^{-4}$
Mean w / mean w^2	0.9999917112 / 0.9999834532
Sample $\widehat{\chi}^2$ / $\widehat{\mathrm{KL}}(\mathsf{Q}\|\mathsf{P})$	3.07238×10^{-8} / 8.30413×10^{-6}
Corrected–uncorrected gradient cosine	0.999999988912
Corrected–uncorrected gradient relative ℓ_2	1.5899435×10^{-4}
Minimum selected 53-bit grid points	15,893,817,923,988
Maximum grid-mass residual	2^{-53}

QMT-S enters the same trainer through a shape-locked $G = 16, H = 4096$ plugin. Its production decoder maintains the original sequence-level jitter as an exact dyadic source cell and commits a token only after outward-rounded containment in the represented CDF is certified. A stable descending-probability top-k prefix, with token IDs breaking ties, is inserted into a zero-padded full-vocabulary FP64 buffer on the model device. Every expansion therefore uses the same-width cumulative-sum, cumulative-maximum, cap, and uniform-floor scan; strict shared prefixes are bitwise stable, and only candidate brackets cross to the host. The decoder preserves logical lanes across EOS and resume and rejects unresolved interval boundaries before committing a token. Independent sampler timing is reported in Table 2. Because QMT-S uses dependent sequence-level sampling, its training estimates are reported as a control and are not covered by Theorem 6.

The matched one-update QMT-S check covered all 16 prompt groups. Every group passed exact lane replay, pathwise containment certification, and prefix-shape invariance. With vocabulary width 151,936, the on-device prefix began at 256 and expanded on demand to an observed maximum of 2,048; candidate-bracket transfer totaled 5,991,360 bytes and full-CDF host transfer remained zero.

B.4 Finite random-source and execution-path audit

The production uniform-plan constructor uses a CPU `torch.Generator` for each derived group seed. It draws FP64 branch decisions and independent jitters, assigns the strata, and constructs the entire $[B, G, H]$ plan on CPU before transfer to the model device. Thus the relevant random source is the CPU stream; GPU checks concern faithful transport and coordinate use. This audit complements the proposal-weight audit, which tests behavior–scoring correspondence after decoding.

Pair frequencies and conditional jitters. We inspect 1,024 independent group-seed streams per sampler at $G = M = 16, H = 4096$, yielding 4,194,304 group positions and 67,108,864 coordinates per arm. Three fixed lane pairs, $(0, 1)$, $(2, 3)$, and $(14, 15)$, are retained separately. The 16-by-16 stratum grid tests each ordered cell against $1/256$. A 32-by-32 grid of complete coordinates tests probability $1/1024$ per cell, and an 8-by-8 grid of reconstructed jitters tests $1/64$ conditional on equal strata. For HO, equal-stratum events are exactly the shared branch. This conditional test detects reused jitter even when stratum-pair frequencies are correct. All distributional checks use complete plans before EOS masking; conditioning on surviving trajectories would change the audit population.

Detection scale and dependence. The unit for uncertainty is a whole group-seed stream, including all its positions. We retain separate tables for each lane pair and cluster by group seed. If C_{bc} is stream b's count for cell c, n_b its relevant sample count, and $N = \sum_b n_b$, we use

$$\widehat{p}_c = \frac{\sum_b C_{bc}}{N}, \qquad \widehat{\mathrm{se}}_c = \frac{\sqrt{\frac{B}{B-1}\sum_b (C_{bc} - \widehat{p}_c n_b)^2}}{N}, \qquad B = 1024. \tag{62}$$

Nominal simultaneous intervals use $\widehat{p}_c \pm z_* \widehat{\mathrm{se}}_c$, with $z_* = \Phi^{-1}(1 - 0.05/(2 \cdot 8064))$. The 8,064 cells cover both samplers, all three lane pairs, and all three grids. Bonferroni adjustment does not require independence between cells or lane pairs. The normal approximation and independence of group-seed streams remain statistical assumptions. Shared-jitter counts vary across streams and use their own denominator.

Table A2: **Finite uniform-pair frequency audit,** $G = M = 16$**.** Each fixed lane pair has 4,194,304 observations from 1,024 whole-group streams, each with 4,096 positions. Entries are maximum absolute cell deviation / largest nominal simultaneous half-width, scaled by 10^5. Equal-stratum jitter uses its conditional sample count. Intervals use group-cluster standard errors and a Bonferroni correction over all 8,064 cells in this table.

Sampler	Lanes	Strata 16^2	Uniforms 32^2	Jitter 8^2
iid	(0,1)	8.25 / 14.65	4.74 / 7.42	52.37 / 114.87
iid	(2,3)	10.75 / 14.53	5.17 / 7.35	67.84 / 113.98
iid	(14,15)	9.75 / 14.54	5.27 / 7.43	56.00 / 115.47
ho	(0,1)	10.23 / 14.76	4.84 / 7.41	73.79 / 116.01
ho	(2,3)	12.09 / 14.65	4.96 / 7.38	55.07 / 114.20
ho	(14,15)	9.80 / 14.79	4.77 / 7.52	80.99 / 114.17

Across the three HO lane pairs, maximum absolute deviations are 1.21×10^{-4} for the stratum grid, 4.96×10^{-5} for the full-coordinate grid, and 8.10×10^{-4} for the conditional-jitter grid. The respective largest simultaneous half-widths are 1.48×10^{-4}, 7.52×10^{-5}, and 1.16×10^{-3}. Every audited cell includes its product-uniform target within its nominal simultaneous interval. The shared branch occurs in 6.2537% of positions, against 6.25% by construction; its marginal group-cluster interval is $[6.2302, 6.2773]\%$. These estimates specify the sampling precision of the coordinate audit.

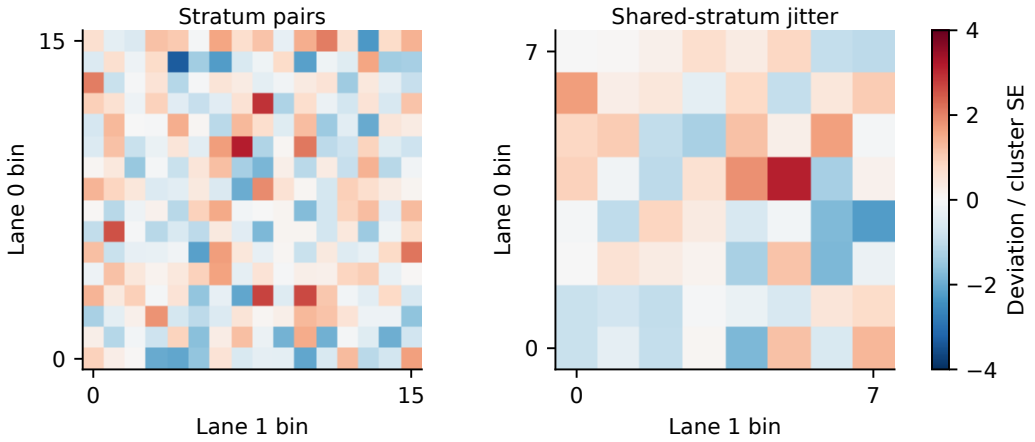


Figure A1: **HO pair-frequency residuals for lanes 0 and 1.** Left: ordered stratum cells. Right: jitter cells conditional on the shared branch. Color gives the signed deviation from the product-uniform cell probability, divided by its group-cluster standard error. These inspect different components of the coordinate pair law.

Position, stream, and execution checks. Table A3 reports complete-coordinate and conditional-jitter correlations, lane-0 lag-one correlations within streams, and lane-0 correlations between 512 disjoint pairs of streams. The disjoint pairing avoids treating overlapping stream pairs as independent clusters. Intervals use a delete-one-cluster jackknife. Table A4 compares FP64 bit patterns through CPU–GPU transfer, inverse lane remapping, and scripted EOS compaction; all comparisons retain original logical group/lane/position identifiers. Seed uniqueness and exact jitter duplication are checked separately. The scripted masks end each lane between positions 2,048 and 4,095; they test coordinate identity, not task-dependent stopping behavior.

Table A3: **Coordinate correlations and cluster uncertainty.** Marginal 95% intervals use a delete-one-cluster jackknife. Pair and lag statistics cluster by the whole group seed; cross-stream statistics use 512 disjoint stream pairs. Correlation checks complement the two-dimensional frequency grids.

Sampler	Statistic	Observations	$\widehat{\rho}$	95% interval
iid	Lanes 0/1	4,194,304	-0.00052	[-0.00148, 0.00045]
iid	Shared-stratum jitter	262,745	-0.00280	[-0.00650, 0.00090]
iid	Lane 0, lag 1	4,193,280	-0.00099	[-0.00197, -0.00002]
iid	Disjoint streams	2,097,152	0.00006	[-0.00132, 0.00144]
ho	Lanes 0/1	4,194,304	-0.00027	[-0.00120, 0.00067]
ho	Shared-stratum jitter	262,300	-0.00261	[-0.00649, 0.00128]
ho	Lane 0, lag 1	4,193,280	-0.00048	[-0.00146, 0.00049]
ho	Disjoint streams	2,097,152	-0.00043	[-0.00181, 0.00095]

Table A4: **Coordinate identity along the execution path.** Equality compares the original FP64 coordinates at the same logical group/lane/position. Scripted EOS masks preserve active coordinates at their original logical indices.

Check	Compared	Mismatches
CPU-to-GPU round-trip bit equality	67,108,864	0
Inverse lane-remap coordinate equality	67,108,864	0
Surviving coordinates under scripted EOS masks	50,358,272	0
Duplicate derived group seeds	1,024	0
Out-of-range or nonfinite HO coordinates	67,108,864	0
Equal jitters in the primary shared pair	262,300	0

These measurements quantify coordinate-level agreement at the listed sample size, grid resolutions, and correlation scales. Appendix A.11 relates finite execution to the mathematical random-source model, and Appendix C.12 summarizes the interpretation of the audit.

B.5 Finite-state checks

Finite-state enumeration computes the exact outcome probabilities of a small categorical model and a two-token autoregressive model by integrating within-bin jitter analytically. The two-token policy has first-token success probability 0.35, second-token probabilities 0.2 and 0.7 conditional on the first token, and XOR terminal reward. The three parameters are the corresponding logits. Its trajectory probabilities are $(0.52, 0.13, 0.105, 0.245)$ and its true objective gradient is $(0.02275, 0.104, -0.0735)$.

Table A5: **Numerical correspondence on finite models.** Values come from enumeration or central finite differences, not language-model runs.

Quantity	Maximum absolute error
Categorical pair law versus product target	2.78×10^{-17}
Two-token trajectory pair law	5.56×10^{-17}
Two-token RLOO mean versus true gradient	8.33×10^{-17}
Negative detached-loss derivative versus true gradient	1.22×10^{-10}
Fixed-baseline REINFORCE covariance versus iid	5.56×10^{-17}

These checks use the idealized, unfloored model with matched behavior and scoring laws, for which $\mathsf{Q} = \mathsf{P}$ and $w = 1$. The implemented uniform support component is treated separately by the proposal correction. Table A1 reports the frozen-batch weight and gradient audit, and Appendix A.11 gives the numerical residual bounds. These checks concern implementation correspondence; final policy quality is evaluated in Section 6.1.

C EXPERIMENTAL PROTOCOL AND SUPPLEMENTARY RESULTS

C.1 DATA, MODELS, AND EVALUATION

This appendix specifies the training, evaluation, and diagnostic protocol. Model identifiers are `Qwen/Qwen3-1.7B` and `Qwen/Qwen3-4B`; every formal record binds its checkpoint, tokenizer, and template revision. Both domains use protocol `qwen3-rlft-prompts-nonthinking-v1`: `tokenize=True`, `add_generation_prompt=True`, and `enable_thinking=False`. Its content ID is `sha256:def0dc7e...46126`, and a pinned-tokenizer golden probe fixes the exact math and code token sequences. The mathematical system message asks for step-by-step reasoning and a boxed final answer; the code message requests one complete Python 3 program in a single Python fence.

The completion horizon was selected without rewards on eight disjoint mathematical prompts, with $G = 4$ and 32 rollouts at each candidate. The no-EOS-at-cap fractions for $H = 1024, 2048, 4096$ were respectively $26/32$, $13/32$, and $2/32$. The immutable rule selected the first fraction at most 0.25, fixing $H = 4096$ before the mechanism comparison or any training run. LoRA applies to attention q/k/v/o and MLP gate/up/down projections, with rank 32, alpha 64, and dropout zero.

Training mathematics comes from `https://huggingface.co/datasets/agentica-org/DeepScaleR-Preview-Dataset`. Training code comes only from the `primeintellect` configuration of `https://huggingface.co/datasets/agentica-org/DeepCoder-Preview-Dataset` at revision `177913a7...ab3db`. The training procedure does not use reference solutions for supervised fine-tuning. The mathematical manifest retains 8,192 training and 512 validation problems after a full-source answer-health and overlap audit. A content-addressed AIME audit pins the complete 30-problem AIME 2024 and AIME 2025 revisions and finds zero normalized-exact matches and zero near matches at the registered 5-token-shingle Jaccard threshold against all 39,055 `math_v3` candidates, the 8,192/512 train/validation manifests, and the frozen P1.4/P1.5 partitions. This leaves the `math_v3` split unchanged. The five-shard DeepCoder audit reads 16,252 source rows and retains all 8,572 unique, supported stdin/stdout problems: 8,060 for training and 512 for validation. Source solutions are neither exported to the training manifest nor executed. The five source hashes, reconstructed manifest, verifier image, reward semantics, sandbox probes, and contamination decision are bound by a separate content-addressed training-readiness decision.

Normalization uses Unicode NFKC and whitespace normalization. The mathematical screen compares against MATH-500 by exact normalized text and exact Jaccard over token 5-grams after an inverted-index candidate search. The code screen compares every retained prompt against a pinned projection of 599 train and 279 test LiveCodeBench-v5 statements; it finds zero exact normalized matches and zero matches at the registered casefolded 5-token-shingle Jaccard threshold of 0.8. This 878-statement projection is the contamination reference, not the final date-window evaluation set. Every exclusion is retained with its reason, and no final-test label selects hyperparameters or checkpoints.

MATH-500 uses the fixed public 500-problem subset; AIME24 and AIME25 each use the complete 30-problem competition set. LiveCodeBench uses the official lightweight execution harness and `release_v5`, restricted to contest dates 2024-08-01 $\leq d <$ 2025-02-01. This window contains 279 problems—175 from AtCoder and 104 from LeetCode—with 9,883 official test cases. We retain the harness commit, Python version, per-test timeout, and extraction configuration with each evaluation record.

Table A6: **Final evaluation settings.** All methods use iid evaluation. Each table cell is an average success probability, not group coverage.

Benchmark	Problems	Samples / problem	Temperature	Token cap
MATH-500	500	4	0.6	8,192
AIME 2024	30	32	0.6	8,192
AIME 2025	30	32	0.6	8,192
LCB-v5 date window	279	10	0.2	8,192

Top-p is 0.95 at final evaluation, identically for all methods. Training uses temperature 1 and no truncation of categorical support. Mathematical verification extracts the final answer and checks numerical or symbolic equivalence under a fixed checker version. The DeepCoder training reward accepts exactly one complete `python` fence, compares ordered whitespace tokens, and equals one only when every selected case is correct. Training and intermediate validation select at most 16 cases per problem using separate content-addressed SHA256 rankings. Each case runs in a fresh process within the pinned Docker image as a non-root user, with a read-only root filesystem, no network, dropped capabilities, and fixed CPU, memory, PID, time, and output limits. Wrong answers, timeouts, runtime errors, output-limit violations, and parse errors receive zero; infrastructure errors fail closed and re-score the stored completion without resampling. The final LiveCodeBench score uses the separately pinned official harness described above.

C.2 Complete training configuration

All comparisons use the same lagged RMS optimizer. Each method is trained under the same three seed identifiers at each model size, with a shared adapter initialization and ordered prompt stream within each seed cohort. Every run uses fresh rollout randomness under its own sampler. There is one update per rollout batch, no trajectory reuse, no momentum, no weight decay, no gradient clipping, and no group-standard-deviation normalization. Full group advantages precede micro-batching; the global trajectory denominator is 256. Every prompt group stays on one of two ranks. FP32 model computation and optimizer state, SDPA, disabled TF32/autocast, and FP64 CDFs match Appendix B.

Table A7: **Shared training and hardware settings.** The same values apply to every arm unless separated by model size.

Parameter	Value
Hardware; ranks	2 A100-80GB GPUs; 2 data-parallel ranks
Prompts / update; responses / prompt	16 (8 math, 8 code); 16
Target budget; checkpoint rule	8M action tokens; last complete update below cap
Learning rate; schedule	3×10^{-5}; constant, no warmup
Lagged RMS v_0, ρ, ϵ	$10^{-4}\mathbf{1},\ 0.99,\ 10^{-8}$
Preconditioner bounds	$[0.01, 100]$
LoRA rank / alpha / dropout	32 / 64 / 0
LoRA modules	attention q/k/v/o; MLP gate/up/down
Training temperature / top-p	1.0 / 1.0 (full support)
Prompt cap: 1.7B / 4B	2,048 / 8,192 tokens
Completion cap; uniform support floor	4,096 tokens; 2^{-28}
Scoring microbatch; target scorer	1; fixed-H, uncached teacher forcing
Reward; KL coefficient	binary terminal verification; $\beta = 0$
HO strata / shared probability	$M = G = 16$ / $1/16$ per token
Matched seeds, every method: 1.7B	7, 11, 23
Matched seeds, every method: 4B	11, 23, 47
Training runs / nominal total budget	18 / 144M action tokens

The support floor gives common finite marginal support; detached trajectory weights are neither clipped nor normalized. Target preservation for iid and HO requires the represented-CDF and random-source premises of Theorem 6. QMT-S shares the scorer and optimizer but retains a dependent pair law. Its outward-enclosure decoder commits only certified tokens and has no approximate containment fallback. These implementation requirements are common to training and the frozen-policy comparisons.

C.3 QMT-S estimator and pairwise support

QMT-S adapts QuasiMoTTo's sequence-level stratified sampler to the common RLOO update in Equation 11. It retains the trajectory proposal weight $w_i = \mathsf{P}(\tau_i)/\mathsf{Q}_i(\tau_i)$ and raw leave-one-out advantages, without a dependence-specific pairwise correction. Here *raw* denotes unstandardized RLOO without that pairwise correction; it does not mean that the trajectory proposal weight is omit-

ted. QuasiMoTTo v1 presents pairwise importance weighting as a conditional theoretical construction and identifies a joint-support obstruction for its lattice sampler (Li et al., 2026, Section 2.4.2). Its GRPO experiments retain dependence-induced group-baseline bias despite using sequence-level off-policy weights (Li et al., 2026, Section 4.2). Our control is a matched RLOO adaptation, with its own model and task protocol.

The corresponding support condition for strict stratification can be derived directly. Fix a prompt and policy and consider ideal continuous coordinates with a common lane-local arithmetic decoder. Let $I_y \subset [0,1)$ be the interval for a complete EOS-terminated or horizon-capped trajectory y, so that $|I_y| = \mathsf{Q}(y)$ under the represented behavior law. For strata $S_k = [k/G, (k+1)/G)$, define $a_{yk} = |I_y \cap S_k|$. A random permutation assigns distinct strata to lanes, and their within-stratum jitters are independent. Thus, for $i \neq j$,

$$\mathsf{Q}_{ij}(y,z) = \frac{G}{G-1} \sum_{k \neq \ell} a_{yk} a_{z\ell} = \frac{G}{G-1} \left[\mathsf{Q}(y)\mathsf{Q}(z) - \sum_{k=0}^{G-1} a_{yk} a_{zk} \right]. \tag{63}$$

The factor combines the ordered stratum-pair probability $1/[G(G-1)]$ with the two conditional densities G. This derivation concerns the stratified control; it does not transfer the lattice-specific joint law to QMT-S.

For the target-scored pair kernel h in Equation 3, direct importance reweighting would use

$$\rho_{ij}(y,z) = \frac{\mathsf{P}(y)\mathsf{P}(z)}{\mathsf{Q}_{ij}(y,z)}. \tag{64}$$

Its change-of-measure guarantee requires proposal support on target-contributing pairs. Two distinct trajectories whose intervals lie wholly inside the same stratum have $\mathsf{Q}_{ij}(y,z) = 0$, even when $\mathsf{P}(y)\mathsf{P}(z) > 0$. Such a missing pair matters when its rewards differ and $h(y,z) \neq 0$; the diagonal case $y = z$ has zero kernel. A positive token-support floor does not remove this joint exclusion. This establishes an obstruction to a general unbiasedness guarantee for direct weighting, without quantifying the omitted contribution or numerical stability in language-model training.

QMT-S therefore measures the behavior of structured sampling under the shared raw RLOO update. Its results characterize that sampling–estimator combination. They neither rank all pairwise-corrected QuasiMoTTo variants nor, by themselves, identify the causal benefit of preserving pairwise independence. The iid–HO comparison varies higher-order dependence within the class of pairwise-independent samplers.

C.4 Endpoint estimates, uncertainty, and seed variation

For a benchmark $\mathcal{D}$ with n_x samples per problem, we report

$$\widehat{A}(\mathcal{D}) = \frac{100}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \frac{1}{n_x} \sum_{i=1}^{n_x} R_{x,i}, \qquad \widehat{A}_{\text{macro}} = \frac{\widehat{A}_{\text{MATH500}} + \widehat{A}_{\text{AIME24}} + \widehat{A}_{\text{AIME25}}}{6} + \frac{\widehat{A}_{\text{LCB}}}{2}. \tag{65}$$

Each endpoint contains 2,000 MATH-500, 960 AIME24, 960 AIME25, and 2,790 LCB responses. Main-table estimates are the arithmetic mean of the three seed-specific scores and their sample SD, with divisor $3 - 1$. The initial-policy score is a shared untrained reference and is not assigned training-seed uncertainty.

For a method contrast, let d_s be the score difference within matched seed s. The training-seed interval is $\bar{d} \pm t_{0.975,2}\, s_d/\sqrt{3}$, where $t_{0.975,2} = 4.30265$, describing run variation on the fixed benchmark population. Separately, 10,000 paired problem bootstrap draws resample problem identities within each benchmark, retain all responses and all three seeds for a selected problem, and condition on the six compared trained checkpoints. Table A8 reports both summaries; their statistical interpretation is given in Appendix C.12.

Table A8: **HO minus iid: distinct uncertainty summaries.** Seed intervals use the three paired training-seed differences and a two-sided Student-t interval with 2 degrees of freedom. Problem intervals use 10,000 paired problem bootstrap draws conditional on all six compared checkpoints. Both are marginal 95% intervals for their respective uncertainty source.

Model	Metric	Mean Δ	Seed interval	Problem interval
1.7B	MATH500	+2.33	[-0.47, 5.14]	[1.45, 3.20]
1.7B	AIME24	+3.51	[0.01, 7.00]	[1.91, 5.21]
1.7B	AIME25	+2.12	[1.33, 2.91]	[0.90, 3.65]
1.7B	LCB	+2.37	[-0.13, 4.86]	[1.70, 3.05]
1.7B	Math mean	+2.65	[0.70, 4.60]	[1.89, 3.44]
1.7B	Macro	+2.51	[0.29, 4.73]	[1.98, 3.01]
4B	MATH500	+1.98	[1.55, 2.42]	[1.27, 2.73]
4B	AIME24	+4.06	[2.62, 5.50]	[2.29, 6.08]
4B	AIME25	+3.02	[2.12, 3.92]	[1.63, 4.83]
4B	LCB	+2.53	[0.29, 4.77]	[1.70, 3.41]
4B	Math mean	+3.02	[2.37, 3.68]	[2.20, 3.92]
4B	Macro	+2.78	[1.64, 3.92]	[2.19, 3.40]

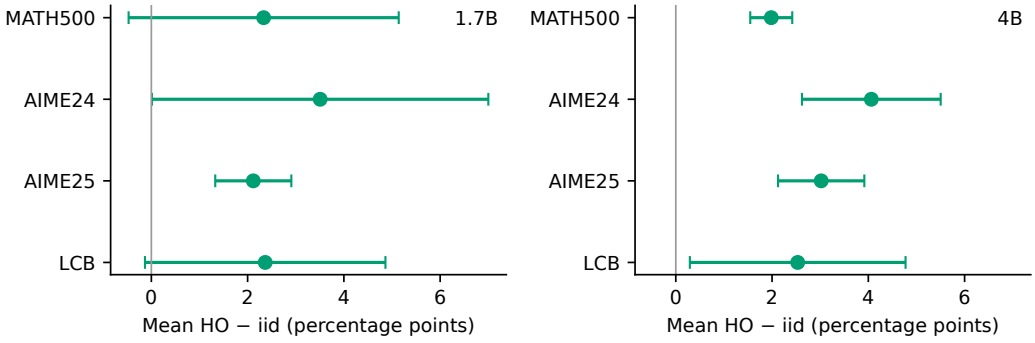


Figure A2: **Per-benchmark endpoint contrasts.** Mean HO minus iid over the three matched seed pairs, with 95% Student-t seed intervals. Conditional problem-bootstrap intervals are reported separately in Table A8.

Table A9: **All matched training-seed results.** Each model has the same three seeds for iid, QMT-S, and HO. Every row gives pass@1 percentages at the selected final checkpoint. The initialization, ordered prompt stream, and evaluation jobs are shared within each seed cohort.

Model	Seed	Sampler	MATH500	AIME24	AIME25	LCB	Macro
1.7B	7	iid-RLOO	76.70	13.12	8.65	16.67	24.75
1.7B	7	QMT-S/RLOO	76.55	13.96	9.17	17.31	25.27
1.7B	7	HO-RLOO	79.15	16.67	11.04	19.10	27.36
1.7B	11	iid-RLOO	77.40	13.65	8.33	17.24	25.18
1.7B	11	QMT-S/RLOO	76.95	14.27	9.58	17.60	25.60
1.7B	11	HO-RLOO	78.55	15.73	10.52	18.57	26.75
1.7B	23	iid-RLOO	76.05	12.19	8.96	16.09	24.25
1.7B	23	QMT-S/RLOO	77.10	13.33	9.27	17.06	25.15
1.7B	23	HO-RLOO	79.45	17.08	10.73	19.43	27.59
4B	11	iid-RLOO	84.05	23.23	16.46	27.99	34.62
4B	11	QMT-S/RLOO	84.45	24.17	17.08	28.60	35.25
4B	11	HO-RLOO	86.20	27.81	19.69	30.65	37.61
4B	23	iid-RLOO	83.55	22.71	16.15	28.32	34.56
4B	23	QMT-S/RLOO	84.00	23.54	16.56	28.85	35.11
4B	23	HO-RLOO	85.55	26.88	18.75	29.89	36.81
4B	47	iid-RLOO	84.70	23.96	16.88	27.53	34.69
4B	47	QMT-S/RLOO	85.10	24.69	17.92	28.42	35.50
4B	47	HO-RLOO	86.50	27.40	20.10	30.90	37.78

C.5 Endpoint training resources

Table A10 reports resources at each selected final checkpoint. Each run uses an eight-million-action-token cap and selects the last complete update at or below it. The table retains actual token counts and numbers of updates, rather than treating the common cap as an identical realized budget.

Table A10: **Per-seed final training resources.** GPU-hours sum committed update time on two A100-80GB GPUs; startup, pauses, evaluation, and overshoot are excluded. Throughput divides action tokens by elapsed update seconds. Peak is the larger device allocation.

Model	Seed	Sampler	Updates	Tokens (M)	GPU-h	Tokens/s	GiB
1.7B	7	iid-RLOO	33	7.980	35.05	126.49	25.55
1.7B	7	QMT-S/RLOO	30	7.659	32.65	130.34	25.35
1.7B	7	HO-RLOO	32	7.910	34.75	126.45	25.55
1.7B	11	iid-RLOO	32	7.862	34.98	124.86	25.62
1.7B	11	QMT-S/RLOO	32	7.791	33.66	128.61	25.46
1.7B	11	HO-RLOO	33	7.816	34.65	125.30	25.59
1.7B	23	iid-RLOO	34	7.922	34.62	127.12	25.49
1.7B	23	QMT-S/RLOO	31	7.731	33.19	129.42	25.40
1.7B	23	HO-RLOO	32	7.887	34.60	126.64	25.53
4B	11	iid-RLOO	30	7.672	67.23	63.40	43.16
4B	11	QMT-S/RLOO	31	7.782	69.96	61.80	43.42
4B	11	HO-RLOO	31	7.840	69.03	63.10	43.18
4B	23	iid-RLOO	31	7.815	69.24	62.70	43.24
4B	23	QMT-S/RLOO	32	7.875	71.44	61.24	43.50
4B	23	HO-RLOO	31	7.754	68.54	62.85	43.27
4B	47	iid-RLOO	30	7.621	66.11	64.05	43.10
4B	47	QMT-S/RLOO	31	7.733	69.15	62.13	43.38
4B	47	HO-RLOO	32	7.797	68.23	63.48	43.15

GPU-hours sum committed update time across the two A100-80GB devices; startup, pauses, evaluation, and overshoot are excluded. Throughput divides actual action tokens by elapsed update seconds, and peak memory is the larger device allocation. The independent sampler profile in Table 2 uses forward-plus-sampler time as its denominator.

C.6 Reward structure during training

A binary-reward group gives zero raw RLOO advantages when all responses fail or all succeed. Table A11 aggregates the token share in these groups over all committed updates. This quantity describes how the generated-token budget is distributed across reward outcomes as the policies evolve.

Table A11: **Token share in zero-advantage groups.** Mean $\pm$ sample SD over three matched training seeds. Within each run, the share weights uniform-reward groups by their generated action tokens.

Model	Sampler	Math (%)	Code (%)
1.7B	iid-RLOO	29.41 ± 0.17	78.95 ± 0.39
1.7B	QMT-S/RLOO	27.35 ± 0.25	76.61 ± 0.13
1.7B	HO-RLOO	22.27 ± 0.25	71.93 ± 0.23
4B	iid-RLOO	24.26 ± 0.12	74.18 ± 0.45
4B	QMT-S/RLOO	22.25 ± 0.31	71.29 ± 0.24
4B	HO-RLOO	17.31 ± 0.33	67.45 ± 0.15

C.7 Frozen-policy measurement protocol

The diagnostic study uses the 1.7B seed-7 cohort: its shared initialization, halfway iid checkpoint, and first four validation prompts per domain in SHA256 order. Each prompt, policy, and sampler receives 16 groups of 16 responses. Separate iid pools contain 64 responses per prompt and checkpoint. The reward matrices comprise 768 groups and 12,288 responses, all with projected-gradient measurements. Reference pools contribute another 1,024 responses. Difficulty is estimated once from these pools and held fixed inside the mechanism bootstrap.

For group success count N_b, coverage is $K^{-1}\sum_b \mathbf{1}\{N_b > 0\}$ and the mixed-group rate is $K^{-1}\sum_b \mathbf{1}\{0 < N_b < G\}$. These are direct frequencies of complete correlated groups. Let ϕ be the final transformer block's adapters and let $A \in \mathbb{R}^{64\times \dim \phi}$ have fixed Rademacher entries scaled by $1/8$. The projected score is $z_i = A\nabla_\phi \log p_\theta(\tau_i \mid x)$; the corrected group estimator is $\widehat{z} = G^{-1}\sum_i w_i A_i z_i$. The projection is identical across conditions and excludes the optimizer preconditioner.

For each prompt, $V = (K-1)^{-1}\sum_b \|\widehat{z}_b - \bar{z}\|_2^2$ and $cV = \bar{c}V$, where $\bar{c}$ is mean generated action tokens per group. Domain and macro summaries average the prompt-specific products. Nested bootstrap intervals use 2,000 draws: prompts are resampled jointly across samplers, then whole groups independently within each sampler. They characterize prompt-and-group sampling variation conditional on the diagnostic checkpoints.

Table A12: **Mechanism by policy and domain.** Coverage and mixed-group rates are percentages over 64 groups. Gradient variance uses all four fixed prompts per domain; no full-parameter variance claim is made.

Policy	Domain	Sampler	Coverage	Mixed	V	$cV/10^3$
initial	math	iid-RLOO	75.00	73.44	15.65	197.75
initial	math	QMT-S/RLOO	78.12	76.56	14.24	180.47
initial	math	HO-RLOO	81.25	78.12	13.50	171.19
initial	code	iid-RLOO	45.31	45.31	6.00	46.00
initial	code	QMT-S/RLOO	50.00	50.00	5.46	42.01
initial	code	HO-RLOO	50.00	50.00	5.52	42.38
halfway	math	iid-RLOO	82.81	81.25	15.28	199.25
halfway	math	QMT-S/RLOO	84.38	82.81	13.91	182.46
halfway	math	HO-RLOO	85.94	84.38	13.10	172.92
halfway	code	iid-RLOO	60.94	60.94	6.60	51.49
halfway	code	QMT-S/RLOO	62.50	62.50	6.01	47.20
halfway	code	HO-RLOO	60.94	60.94	5.92	46.90

Table A13: **Difficulty-stratified diagnostics.** Strata use the fixed 64-response iid pool: sparse $\widehat{p} \le 1/16$, intermediate $1/16 < \widehat{p} \le 1/2$, dense $\widehat{p} > 1/2$. Coverage is a percentage. The dense stratum contains one prompt.

Policy	Stratum	Prompts	Coverage iid/HO	V iid/HO
initial	sparse	5	37.50 / 46.25	4.64 / 3.15
initial	intermediate	2	96.88 / 96.88	20.00 / 17.52
initial	dense	1	100.00 / 100.00	23.40 / 25.27
halfway	sparse	3	39.58 / 41.67	2.76 / 1.81
halfway	intermediate	4	89.06 / 90.62	14.25 / 11.65
halfway	dense	1	100.00 / 100.00	22.26 / 24.04

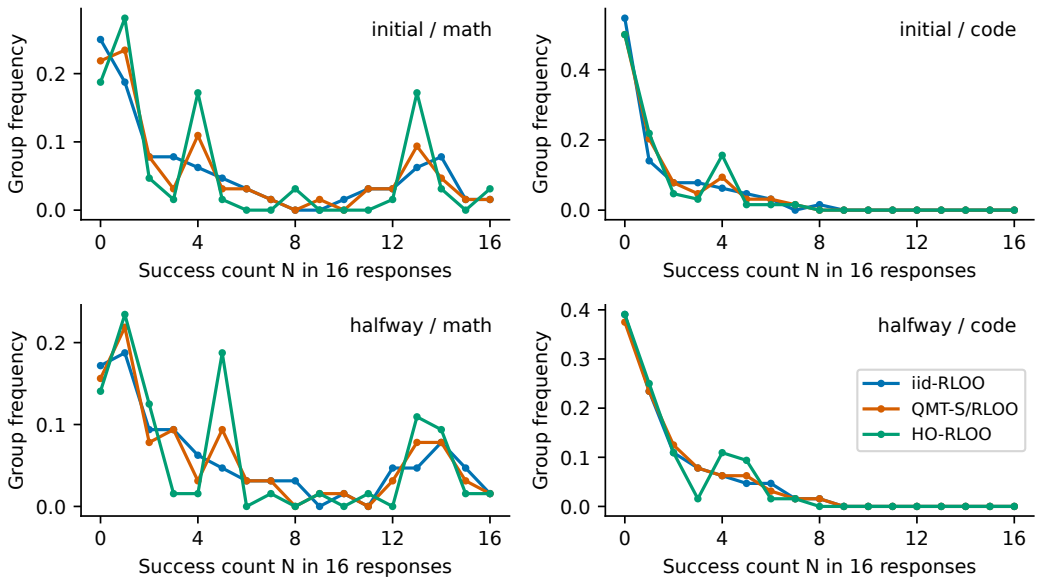


Figure A3: **Distribution of successes per group.** Each curve pools 64 groups at one policy and domain and includes both zero and all-success outcomes. The histograms show how the samplers distribute reward contrasts across complete groups.

C.8 Strata ratio at fixed rollout count

The choice $M = G$ makes every distinct-branch group cover all strata once and introduces no additional stratum-count tuning. It also uses the largest shared-branch probability within this family. Increasing M changes higher-order structure while retaining the pair-law calibration. For three distinct lanes and any $M \geq G \geq 3$,

$$\mathbb{P}_{\mathrm{HO}}(Z_1 = Z_2 = Z_3) = \frac{1}{M}, \qquad \mathbb{P}_{\mathrm{iid}}(Z_1 = Z_2 = Z_3) = \frac{1}{M^2}, \\ \mathbb{P}_{\mathrm{HO}}(Z_i = a, Z_j = b) = \frac{1}{M^2}. \tag{66}$$

The shared-branch probability must remain $1/M$ when M changes; changing it independently would invalidate the last identity.

Relation to the iid limit. For fixed G, the ideal uniform-group law converges to iid as $M \to \infty$. To see this, let D be the event of all distinct strata and S that of all equal strata. Under iid, their probabilities are $(M)_G/M^G$ and M^{1-G}, where $(M)_G = M(M-1)\cdots(M-G+1)$. HO assigns mass $1 - 1/M$ to D and $1/M$ to S, uniformly within each event, with identical independent conditional jitters. For $G \geq 2$, the HO density is at least the iid density on both events and vanishes on other collision patterns. Therefore

$$\mathrm{TV}(\mathcal{U}_{\mathrm{HO}}^{G,M}, \mathcal{U}_{\mathrm{iid}}^{G}) = 1 - \frac{(M)_G}{M^G} - M^{1-G} \leq \min\left\{1, \frac{\binom{G}{2}}{M}\right\}. \tag{67}$$

The bound follows from the iid collision union bound and quantifies the joint law of one uniform group. At finite M, reward coverage and gradient variance also depend on how the finer strata align with reward-bearing trajectories.

Frozen-policy comparison. We fix $G = 16$ and compare iid with HO at $M \in \{16, 32, 64, 128\}$ using the 1.7B seed-7 initialization and halfway iid checkpoint. Every condition uses the same eight prompts, numerical scorer, 64-dimensional projection, verifier, and 64 groups per prompt. Each condition has a matched budget of 8,192 responses per checkpoint; realized token counts enter cV. This study adds 5,120 groups and 81,920 responses through fresh sampling at the fixed policies,

complementing the smaller matrix in Appendix C.7. All reported V values center within prompt and sampler; cV averages prompt-specific products of token cost and projected variance.

Table A14: **Varying strata at fixed rollout count, $G = 16$.** Each condition uses eight fixed prompts and 64 groups per prompt at the same policy, for 8,192 responses. Coverage, mixed rate, and reward are percentages. V averages within-prompt projected variance; cV averages prompt-specific token-cost products. Ratios use the iid row at the same checkpoint.

Policy	Sampler / M/G	Coverage	Mixed	Reward	V	cV ratio
initial	iid	60.94	60.55	18.55	10.84	1.000
initial	HO / 1	66.02	65.23	18.53	9.51	0.926
initial	HO / 2	66.80	65.82	18.53	10.07	0.963
initial	HO / 4	63.48	62.89	18.43	10.47	0.973
initial	HO / 8	61.72	61.33	18.59	10.73	0.972
halfway	iid	71.88	71.09	21.45	10.96	1.000
halfway	HO / 1	75.00	74.02	21.33	9.51	0.914
halfway	HO / 2	76.37	75.20	21.52	10.08	0.956
halfway	HO / 4	73.44	72.66	21.42	10.52	0.984
halfway	HO / 8	72.27	71.48	21.47	10.83	1.010

The default ratio gives projected cV ratios of 0.926 at initialization and 0.914 at the halfway checkpoint. Increasing M/G to two raises coverage from 66.02% to 66.80% and from 75.00% to 76.37%, respectively, while the cV ratios rise to 0.963 and 0.956. Ratios four and eight move these reward and gradient summaries toward the iid reference. Substantial joint uniform dependence remains at these finite ratios: Equation 67 gives total variation approximately 0.624 at $G = 16, M = 128$. The reward-level trends thus reflect both the joint sampling law and its interaction with the decoder and verifier.

Figure A4 summarizes the coverage and cost–variance tradeoff; Table A14 gives the complete measurements.

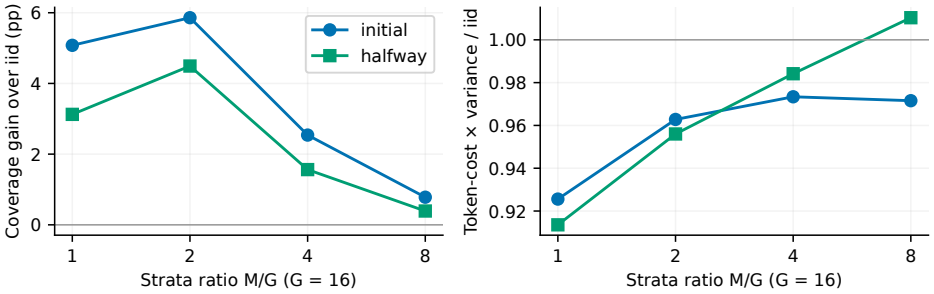


Figure A4: **Changing strata while keeping $G = 16$.** Point estimates from 64 groups per prompt on the fixed eight-prompt panel. Left: coverage difference from iid. Right: projected token-cost–variance relative to iid. The default $M/G = 1$ has the lowest cost–variance, whereas $M/G = 2$ gives the highest coverage among the tested settings.

The comparison supports $M = G$ as a simple default with a favorable cost–variance profile on this panel, while a larger ratio achieves higher coverage. The number of rollout lanes stays fixed here; changing G together with $M = G$, as in Appendix C.9, also changes the response budget. The two studies address different design choices. Execution overhead additionally depends on the CPU permutation size, which grows with M.

C.9 Group size and estimator ablations

The structural study reuses the initial size-16 groups in Appendix C.7 and adds size 8 and 32 for iid and HO, with 16 groups per prompt, plus an all-distinct size-16 condition. These additions comprise 640 groups and 12,288 responses. This study and the original frozen-policy matrix contain 1,408 comparison groups and 24,576 responses, plus 1,024 reference responses. The separate fixed-G

strata-ratio study adds 81,920 responses, giving 107,520 diagnostic and reference responses in total. The uniform audit generates coordinates only. No additional policy training is needed; standardization reuses stored HO rewards and projected gradients.

Table A15: **Group-size ablation at initialization.** Each condition has 64 groups per domain. cV accounts for the increase in cost with group size; reward and gradient summaries use the same fixed prompts.

Domain	G	Sampler	Coverage	Mixed	V	$cV/10^3$
math	8	iid-RLOO	64.06	59.38	33.49	209.29
math	8	HO-RLOO	65.62	60.94	28.68	180.78
math	16	iid-RLOO	75.00	73.44	15.65	197.75
math	16	HO-RLOO	81.25	78.12	13.50	171.19
math	32	iid-RLOO	87.50	87.50	8.57	215.88
math	32	HO-RLOO	93.75	92.19	7.34	188.94
code	8	iid-RLOO	37.50	37.50	12.95	49.85
code	8	HO-RLOO	35.94	35.94	11.86	45.33
code	16	iid-RLOO	45.31	45.31	6.00	46.00
code	16	HO-RLOO	50.00	50.00	5.52	42.38
code	32	iid-RLOO	56.25	56.25	3.30	50.38
code	32	HO-RLOO	57.81	57.81	3.05	46.42

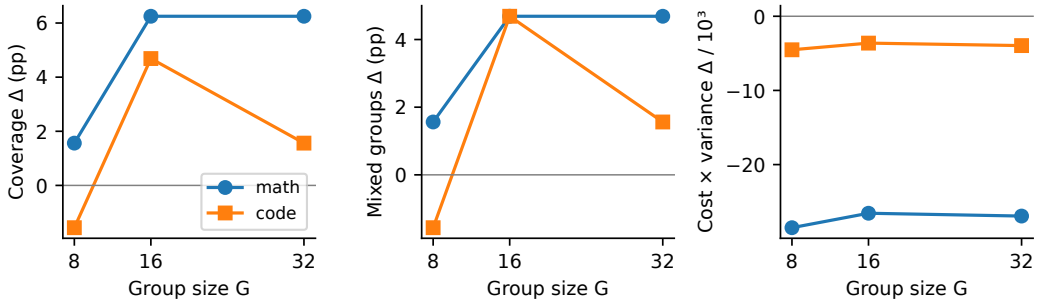


Figure A5: **HO minus iid across group sizes.** Point estimates at initialization, with identical prompt panels and 16 groups per condition. The panels compare coverage, mixed-group rate, and projected cost–variance, charging for every generated token.

For the reference-gradient diagnostic, let δ_p be the difference between a method's prompt-mean projected estimator and its independent 64-response iid reference mean. We report

$$d = \left(\frac{1}{8}\sum_p \|\delta_p\|_2^2\right)^{1/2}, \qquad s = \left(\frac{1}{8}\sum_p \operatorname{tr}\left(\widehat{\Sigma}_{g,p}/16 + \widehat{\Sigma}_{\mathrm{ref},p}/64\right)\right)^{1/2}. \tag{68}$$

The reference estimator is the mean of proposal-weighted reward scores. Standardized advantages use $(R_i-\bar{R})/(\sqrt{\bar{R}(1-\bar{R})}+10^{-6})$. For binary rewards, its projected gradient is obtained from raw RLOO by multiplying by $(G-1)/[G(\sqrt{\bar{R}(1-\bar{R})}+10^{-6})]$. Uniform-reward groups remain zero. We therefore report each estimator's variance together with its mean-gap magnitude and reference uncertainty.

Table A16: **Estimator and branch ablations.** All-distinct removes the shared branch. Standardization is recomputed on the same HO groups. d is the RMS distance of prompt means from the independent reference gradient; s combines group and reference sampling scales.

Design	Coverage	V	d	s
iid-RLOO + raw	60.16	10.83	0.09	0.96
HO-RLOO + raw	65.62	9.51	0.08	0.91
All-distinct + raw	67.19	7.47	0.91	0.84
HO + standardized	65.62	67.60	0.83	2.11

C.10 DIVERSITY AND NUMERICAL CONDITIONING

Table 4 in Section 6.3 reports sequence diversity and collision statistics. The additional descriptors below use the same frozen-policy groups and distinguish token uniqueness, entropy, prefix overlap, and n-gram diversity.

Table A17: **Additional diversity descriptors.** Token uniqueness divides distinct next tokens by active lanes. Entropy is divided by the logarithm of the active-lane count. Duplicate fraction is one minus sequence uniqueness. Prefix length is the pair-mean longest common prefix divided by the shorter response length. Distinct-n divides unique n-grams by total n-grams within a group.

Policy	Sampler	Token unique	Entropy	Duplicates	Prefix	D-1	D-3
initial	iid-RLOO	0.629	0.775	0.050	0.066	0.371	0.810
initial	QMT-S/RLOO	0.628	0.776	0.057	0.067	0.369	0.810
initial	HO-RLOO	0.652	0.784	0.044	0.062	0.383	0.827
halfway	iid-RLOO	0.631	0.775	0.057	0.066	0.370	0.809
halfway	QMT-S/RLOO	0.628	0.773	0.051	0.065	0.369	0.810
halfway	HO-RLOO	0.654	0.783	0.038	0.062	0.381	0.828

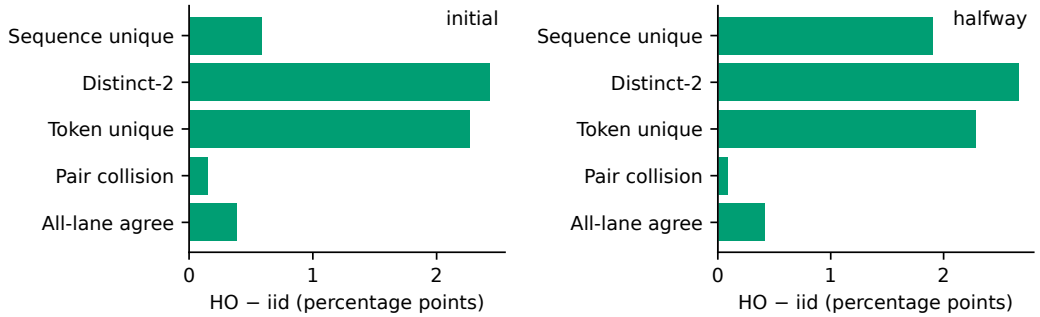


Figure A6: **Changes in rollout diversity.** Group-averaged descriptors from the same size-16 diagnostic groups. Greater sequence and token uniqueness coexists with nearly unchanged pair-token collision and a small increase in all-lane agreement.

Table A18: **Proposal-weight conditioning in the frozen-policy study.** Values pool both domains at $G = 16$. nESS is divided by group size. All groups have zero support, underflow, and nonfinite violations. These measurements characterize numerical conditioning on the diagnostic panel.

Policy	Sampler	Max $\lvert \log w \rvert$	Min nESS
initial	iid-RLOO	0.000713	0.999999927
initial	QMT-S/RLOO	0.000621	0.999999921
initial	HO-RLOO	0.000608	0.999999938
halfway	iid-RLOO	0.000644	0.999999926
halfway	QMT-S/RLOO	0.000681	0.999999929
halfway	HO-RLOO	0.000670	0.999999938

The small logged proposal discrepancies and near-unit nESS indicate well-conditioned, unclipped proposal weights on this panel. Appendix A.11 relates the numerical diagnostics to the residual terms in the implementation model.

C.11 PUBLISHED COMPARISON CONTEXT

Table A19 places the study alongside adaptive rollout allocation and prefix reuse. Each panel retains its source's evaluation and cost definitions. Our controlled sampler comparison is reported in Table 1.

Table A19: **Source-reported context.** Accuracy/pass@1 in percent. Panel (a): PAIR v2, Table 1 (Nomand et al., 2026), a 2026 preprint. Panel (b): PROS, Table 2 (Huang & Wan, 2026), ICLR 2026. All backbones are Qwen3; costs use the indicated source-specific units.

(a) PAIR: MATH/TACO training tracks

Model	Method	MATH500	AIME24	LiveCode	Token ratio
1.7B	GRPO	81.2	24.7	18.6	1.00
1.7B	PAIR	84.8	29.4	22.1	0.49
4B	GRPO	86.4	33.1	26.4	1.00
4B	PAIR	89.9	38.5	30.8	0.48

(b) PROS: AIME-Old training

Model	Method	AIME24	AMC23	Minutes/iteration
4B	PPO	21.25	52.59	5.83
4B	PPO + PROS	27.40	62.12	6.21

PAIR reports separate MATH and TACO training tracks, $G = 16$, an 8,192-token response cap, and BF16 execution. Its token denominator is full-group GRPO at the same model scale; LiveCode denotes its post-cutoff subset. PROS uses AIME problems from before 2024 for panel (b). Its cost unit is average per-iteration GPU wall-clock in minutes. The source's general setup uses $G = 8$ and a 6,144-token response cap. These panels provide context for the matched local experiments; model revisions and numerical evaluation-decoding settings are not reported (NR) in the source registry.

C.12 Scope and statistical interpretation

Evaluation covers two Qwen3 sizes, mathematics and code under the stated non-thinking protocol, and an eight-million-token training budget. Three matched seeds give Student-t intervals with two degrees of freedom under approximately Gaussian seed differences; problem-bootstrap intervals condition on the evaluated checkpoints. These intervals describe separate uncertainty sources and are neither joint nor multiplicity-adjusted. AIME has 30 problems per year, so repeated completions reduce decoding noise within a fixed population. Broader model families, tasks, and training scales remain directions for evaluation.

Frozen-policy diagnostics use eight prompts from the 1.7B seed-7 cohort and a 64-dimensional unpreconditioned projection of the final block's adapter scores. They measure group reward structure and subspace variance, not full-parameter preconditioned variance or a causal explanation for benchmark gains. The decompositions retain heterogeneous outcomes: halfway code coverage is unchanged, the dense initial mathematics prompt has higher HO variance, and HO code coverage at $G = 8$ is 1.56 points below iid (Tables A12, A13, and A15). The lowest projected cV at $M = G$ is a result within the tested strata settings, not a universal optimum. Theorem 2 and Appendices A.6–A.7 characterize the improvement boundary, reversed-variance regime, and neutral geometry.

Resource reporting distinguishes actual endpoint training tokens, committed-update GPU time, and independent sampler service time. Endpoint token counts vary with indivisible updates. Sampler and zero-advantage token shares describe components of resource use, not equivalent fractions of total training time. Theorem 7 concerns one update; benchmark comparisons assess final quality.

The finite-source audit pertains to the listed pairs, seeds, grids, and execution paths. Agreement does not certify exact PRNG independence, unseen configurations, or a tight trajectory-distribution bound (Appendix A.11). Small gradient-mean gaps and near-matching collision rates are descriptive, not equivalence tests. Standardization evaluates stored-group estimators rather than a full GRPO system. Appendix C.3 specifies the QMT-S raw-RLOO control, its pairwise-correction scope, and the joint-support condition.

D EXTENDED DESIGN DISCUSSION

Preserved moments and group-level variation. At a frozen policy, a marginal-preserving sampler leaves the expected reward of an individual rollout unchanged. For binary rewards, pairwise independence also fixes $\mathbb{E}N = Gp$ and $\mathrm{Var}(N) = Gp(1-p)$, while higher-order dependence changes the distribution of discoveries across a finite group. In the aligned sparse example, the frequent distinct branch provides reward contrasts and rare shared groups preserve the required moments. This organizes the policy's existing capability into group-level learning signals.

Separating sampling from the learning target. Arbitrary pairwise repulsion, nonlinear group normalization, and changed rewards can all change the mean update. Raw RLOO provides an analytically tractable setting to isolate higher-order sampling effects at a fixed expected gradient. GRPO (Shao et al., 2024) is also a group estimator, but its advantage normalization introduces nonlinear coupling: the sampler's pair law remains intact, yet no longer guarantees mean preservation (Appendix A.12). Extending higher-order design to normalized estimators requires accounting for changes in both mean and variance.

The estimator-order principle in Appendix A.5 makes the distinction explicit: an order-r kernel needs only r-way laws to determine its mean, while its covariance can depend on laws of up to $2r$ members. For a fixed-baseline REINFORCE average, $r = 1$ and pairwise independence already fixes the variance. The random group baseline is what makes RLOO responsive to higher-order design.

Coverage and estimator variance. The Bernoulli calculation uses the same success count to describe two quantities: $\mathbf{1}\{N > 0\}$ for coverage and $N(G-N)/[G(G-1)]$ for the update. An all-success group contributes to coverage and has zero RLOO contrast. The exact variance regimes in Appendix A.6 make the distinction explicit, motivating separate measurement of discovery frequency and estimator variability.

Transfer to autoregressive trajectories. The trajectory-pair theorem applies across reward geometries, whereas the scalar benefit uses the alignment assumption. In an autoregressive decoder, prefixes change token CDFs and a verifier evaluates the complete response. Tokenwise and sequence-coordinate constructions can therefore have different higher-order reward geometry despite agreeing on every marginal. The theory identifies the preserved pair law and the conditions for variance improvement; numerical diagnostics examine the correspondence between the normalized target and its finite implementation.

Extensions allowed by the constraint. The construction permits any integer $M \geq G$. It also permits a mixture of feasible samplers chosen using earlier history: conditional on that choice, each pair law is the same product, so averaging over the choice preserves it. This gives a mathematical entry point for adapting higher-order structure to reward geometry while keeping the current group's outcomes separate from sampler selection. The training method fixes $M = G$; Appendix C.8 examines alternative fixed strata counts, and Appendix A.10 derives an explicit trajectory-KL extension of the objective.