SHAPLEY-GUIDED ROLLOUT SELECTION FOR REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards improves language-model reasoning, but reducing training costs through response selection risks losing useful learning signals that reward statistics alone cannot identify. We aim to select a small subset of generated responses that preserves the complete group's learning direction under a fixed update budget. We introduce Shapley Selection Rollouts (SSL), which values each response by its average marginal contribution to recovering the group's direction across different response subsets. SSL measures this contribution using compact token-score representations as proxies for update directions, then passes the highest-valued responses to the existing policy optimizer without changing its objective. Across three language models and seven mathematics benchmarks, including GSM8K, MATH500, and AIME24/25, SSL improves mean benchmark accuracy by 0.49–0.62 percentage points over the strongest evaluated alternative when selecting 16 of 64 responses per prompt. By valuing responses in relation to one another, SSL provides a practical approach to preserving useful learning signals within a limited policy-update budget.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves language-model reasoning by learning from automatically checked solutions, making it attractive for mathematics and code generation (Xu et al., 2025). Group Relative Policy Optimization (GRPO) makes such training more accessible by deriving learning signals from groups of responses without a separate critic (Shao et al., 2024). However, updating the policy on every generated response remains memory- and compute-intensive, creating a need to retain the learning value of broad exploration within a smaller update budget.

Rollout down-sampling addresses this bottleneck by generating a larger candidate pool and updating on a selected subset. PODS shows that preserving reward diversity can improve the accuracy–cost trade-off (Xu et al., 2025), while CPPO reduces update work by pruning responses with small absolute advantages (Lin et al., 2025). These results show the value of selective updates. Their effect on learning also depends on update direction: responses with similar rewards can induce different parameter changes, and their contributions can reinforce or cancel one another. Thus two subsets with the same reward distribution can preserve different learning directions.

Recent work recognizes important parts of this difficulty. SALT addresses cancellation by reweighting group-relative learning signals (Chang et al., 2026), and GeoAlign improves stability by curating responses for directional consistency (Zhou et al., 2026). These methods address conflict and consensus in updates. Preserving a complete group's learning signal under a fixed subset budget further requires evaluating responses jointly. This is a context-dependent problem: a response may repeat information already retained or supply a contribution that would otherwise be lost. Moreover, removing responses can change the relative rewards of those that remain; D3S explicitly addresses this estimation issue by retaining full-group advantages (Wang et al., 2026). A central gap therefore remains between identifying individually useful responses and identifying responses whose joint contribution preserves the learning signal available before down-sampling.

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.

> **Research question.** How can we select a small subset of rollouts that preserves the full group's learning direction under a fixed update budget?

Our guiding principle is to value each rollout by how much it helps recover the complete group's update direction across different subset contexts. We introduce *Shapley Selection Rollouts* (SSL), which implements this principle through average marginal-contribution attribution (Shapley, 1953). A compact representation of token scores makes these comparisons practical while accounting for changes in group-relative rewards, and the highest-valued rollouts are passed to the unchanged policy optimizer. We assess this proxy-based selection through independent parameter-gradient diagnostics and paired normalization controls, alongside downstream accuracy and computational cost.

Our contributions are threefold:

- We characterize when reward-based selection misses useful update directions, finding the largest fidelity gains in groups with balanced rewards and strong directional conflict. A conditional analysis connects update fidelity to retained local optimization progress.
- SSL provides context-dependent rollout valuation within each prompt, enabling budgeted selection without an auxiliary learned selector or changes to the policy objective. Paired controls distinguish the effect of selected responses from the effect of advantage normalization.
- Across three backbones and seven mathematics benchmarks, SSL improves the mean score over the strongest evaluated alternative by 0.49–0.62 percentage points in the fixed-pool comparison. On Qwen3-4B-Base, it also achieves the highest mean score among evaluated systems under each of three separately matched resource budgets.

2 Related Work

Budget-aware rollout selection. Within GRPO (Shao et al., 2024), reward-based pruning prioritizes advantage magnitude in CPPO (Lin et al., 2025) and reward variance in PODS (Xu et al., 2025). D3S extends selection across groups and tokens while retaining full-group advantages (Wang et al., 2026). Beyond post-generation filtering, contextual scheduling and replay in CBS (Lu et al., 2026), online pruning in ARRoL (Xu et al., 2026), and predictive allocation in VIP (Nguyen et al., 2026) address broader compute bottlenecks. These strategies act at different stages of the rollout pipeline, so their costs depend on both how many responses are generated and how many enter the update. Directional coverage adds information about how signed rollout contributions combine. Evaluating this aspect of selection requires holding the generated pool fixed. SSL selects completed rollouts within each prompt using directional fidelity and separately controls advantage normalization.

Update geometry and subset approximation. GRAD-MATCH constructs weighted subsets to match training or validation gradients (Killamsetty et al., 2021); GradAlign uses validation-gradient alignment to select RL training prompts (Yang et al., 2026). For group-based RLVR, SALT reweights advantages to mitigate signed gradient cancellation (Chang et al., 2026), while GeoAlign curates responses using learned latent-direction consensus (Zhou et al., 2026). SSL attributes marginal directional contributions across rollout coalitions with recomputed reward statistics. Recomputing these statistics couples a response's value to the other members of its subset: adding one rollout can change the signed contributions of those already present. It evaluates alignment in a fixed token-score representation, requiring no learned projector or validation gradients, and assesses parameter-gradient fidelity independently.

Shapley valuation of training contributions. Data Shapley applies cooperative-game attribution (Shapley, 1953) to training-data value, but evaluating performance-based utilities can require repeated model fitting (Ghorbani & Zou, 2019). Parallel Shapley uses a generative reward model to assign rewards to paths within a parallel reasoning response (Zhang et al., 2026). For budgeted selection of complete rollouts, SSL defines a prompt-local alignment game with coalition-dependent reward centering. Cached token-score features enable exact or permutation-based valuation without coalition-wise retraining or an auxiliary reward model. Averaging marginal contributions across subset contexts accounts for how a rollout's signed direction overlaps with, or complements, directions already represented in a candidate subset.

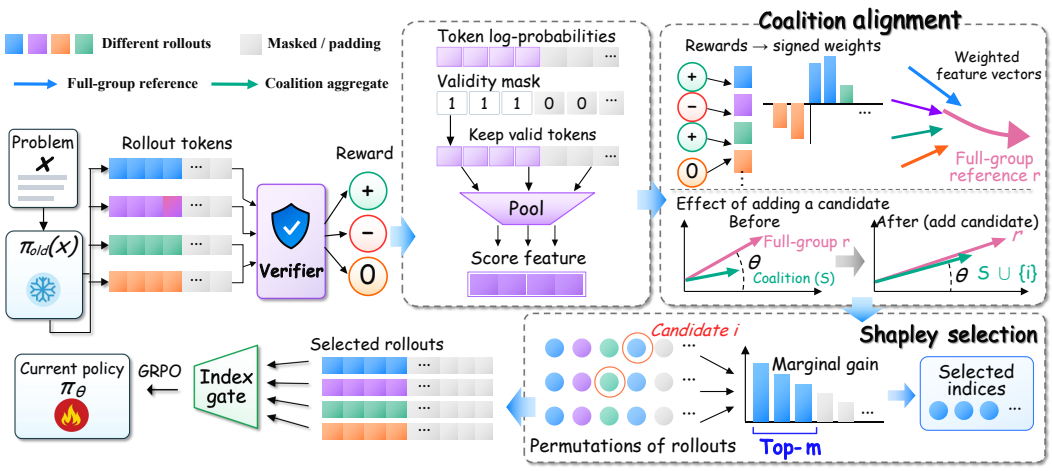

Figure 1: **Overview of Shapley Selection Rollouts (SSL).** A frozen policy generates rollouts for one prompt, and a verifier assigns rewards. Token log-probabilities are masked, length-normalized, and linearly pooled into score features (shown for one rollout). Coalition-centered rewards weight the features to form an aggregate, whose cosine alignment with the fixed full-group reference defines the coalition utility. Averaging marginal utility changes over sampled permutations estimates Shapley values; small groups use exact evaluation. The top-m rollouts pass through the index gate to the unchanged GRPO update.

3 Method

3.1 Overview

For each prompt, RLVR generates more rollouts than the update budget can accommodate. We introduce *Shapley Selection Rollouts* (SSL) to select a smaller group while preserving useful update directions, following the down-sampling setting of Xu et al. (2025). SSL encodes each rollout as a compact token-score vector, then compares the advantage-weighted aggregates of each subset and the full group by cosine similarity. A rollout's Shapley value (Shapley, 1953) is its average marginal contribution to this score across rollout orderings. SSL retains the highest-valued rollouts and passes them to the unchanged GRPO objective. Figure 1 shows the workflow.

3.2 Prompt-local rollout group

For a prompt x, a fixed policy snapshot generates n rollouts, of which m will enter the update:

$$\mathcal{R}(x) = \{(o_i, r_i, \ell_i, m_i)\}_{i \in N}, \qquad N = \{1, \dots, n\}, \qquad 1 \le m < n. \tag{1}$$

Here N indexes the full group and m is the update budget. Rollout $o_i = (y_{i,1}, \dots, y_{i,L_i})$ has verifier reward $r_i \in \mathbb{R}$, padded old-policy token log-probabilities $\ell_i \in \mathbb{R}^T$, and valid-token mask $m_i \in \{0,1\}^T$. Each rollout has $L_i \ge 1$ valid response tokens; padding and verifier-only positions are masked out. Selection uses only the current group and fixed hyperparameters.

3.3 Linear score representation

Storing a parameter-gradient vector for every rollout is costly. SSL compresses token log-probabilities into a fixed-dimensional score vector. With $\mathbf{1}$ denoting the all-ones vector and $L_i = \mathbf{1}^\top m_i$, define

$$q_i = \frac{m_i \odot \ell_i}{\max(1, L_i)}, \qquad z_i = P q_i \in \mathbb{R}^d, \qquad Z = \begin{bmatrix} z_1^\top \\ \vdots \\ z_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad P \in \mathbb{R}^{d \times T}. \tag{2}$$

Masking and length normalization give q_i; the fixed pooling matrix P maps it to z_i, which forms row i of Z. These score vectors define the representation used for Shapley valuation; parameter-gradient fidelity is assessed separately. Appendix A.1 specifies the masking and pooling conventions.

3.4 COALITION PROXY AND UTILITY

To capture the signed direction contributed by a rollout, SSL evaluates each subset, or *coalition*, by how well its combined representation aligns with the full group. For a nonempty coalition $S \subseteq N$, define

$$\mu_S = \frac{1}{|S|}\sum_{i\in S} r_i, \qquad \sigma_S = \sqrt{\frac{1}{|S|}\sum_{i\in S}(r_i-\mu_S)^2}, \quad a_i(S) = \frac{r_i-\mu_S}{\sigma_S+\varepsilon_{\text{adv}}},$$
$$[\mathbf{a}(S)]_i = \begin{cases} a_i(S), & i \in S, \\ 0, & i \notin S, \end{cases} \quad \widetilde{g}(S) = \frac{1}{|S|} Z^\top \mathbf{a}(S), \qquad \widetilde{g}_N = \widetilde{g}(N). \tag{3}$$

Here μ_S and σ_S are the coalition's reward mean and population standard deviation. The GRPO-style advantages $a_i(S)$ form the zero-padded vector $\mathbf{a}(S) \in \mathbb{R}^n$, with fixed stabilizer $\varepsilon_{\text{adv}} > 0$. The full-group reference $\widetilde{g}_N$ stays fixed throughout selection. The positive factors $|S|^{-1}$ and $(\sigma_S + \varepsilon_{\text{adv}})^{-1}$ match the update weighting but cancel within the cosine ratio. The coalition mean determines the signed combination of score vectors.

We measure directional fidelity with the coalition utility

$$v(S) = \begin{cases} \dfrac{\widetilde{g}(S)^\top \widetilde{g}_N}{\|\widetilde{g}(S)\|_2 \, \|\widetilde{g}_N\|_2}, & S \neq \varnothing,\ \|\widetilde{g}(S)\|_2 > \varepsilon_{\cos},\ \|\widetilde{g}_N\|_2 > \varepsilon_{\cos}, \\ 0, & \text{otherwise}, \end{cases} \qquad \varepsilon_{\cos} > 0. \tag{4}$$

Thus $v(\varnothing) = 0$ and $v(S) \in [-1, 1]$. The separate threshold $\varepsilon_{\cos}$ handles zero or negligible directions, whereas ε_{adv} stabilizes reward normalization. This utility measures fidelity in score space; verifier accuracy and parameter-gradient fidelity are evaluated independently.

3.5 SHAPLEY VALUATION AND SELECTION

Adding a rollout changes both the reward center and the signed score aggregate. Its Shapley value (Shapley, 1953) averages the resulting utility change over all rollout orderings. We compute this value exactly for small groups and estimate it from sampled permutations for larger groups:

$$\phi_i = \sum_{S \subseteq N\setminus\{i\}} \frac{|S|!\,(n-|S|-1)!}{n!}\big[v(S\cup\{i\}) - v(S)\big],$$
$$\widetilde{\phi}_i = \begin{cases} \phi_i, & n \le n_{\text{exact}}, \\ \dfrac{1}{K}\displaystyle\sum_{k=1}^{K}\big[v(B_i^{(k)}\cup\{i\}) - v(B_i^{(k)})\big], & n > n_{\text{exact}}, \end{cases} \qquad S_m = \text{Top}_m\{\widetilde{\phi}_i : i \in N\}. \tag{5}$$

Here ϕ_i is the exact attribution, n_{exact} the enumeration threshold, and K the permutation budget. The set $B_i^{(k)}$ contains the rollouts preceding i in uniformly sampled permutation k, making the estimator unbiased for ϕ_i. The operator Top_m returns the indices of the m largest values, breaking ties by increasing rollout index. This rule ranks individual marginal contributions; maximizing $v(S)$ over size-m coalitions is a separate combinatorial objective.

Exact enumeration evaluates each of the 2^n coalition masks once, then accumulates marginal differences. Cached score aggregates give $\mathcal{O}(2^n(d+n))$ time. Storing all coalition vectors requires $\mathcal{O}(2^n d)$ memory; retaining scalar utilities and the needed aggregate sums reduces storage. Permutation sampling updates one prefix coalition per step, taking $\mathcal{O}(Knd)$ time. We use $n_{\text{exact}} = 16$ and $K = 512$. Appendix A.2 details caching and estimator-variance accounting.

3.6 SELECTION VERSUS ADVANTAGE NORMALIZATION

A selected group normally recomputes its reward statistics. To distinguish this effect from the choice of rollouts, we also evaluate the same S_m using full-group advantages:

$$\widetilde{g}_{\text{fix}}(S_m) = \frac{1}{|S_m|}\sum_{i\in S_m} a_i(N) z_i. \tag{6}$$

The control in equation 6 changes only update evaluation; the frozen rollout group, attributions, and selected indices are shared across both conditions. Appendix A.3 details the paired comparison.

3.7 GRPO UPDATE

SSL passes the selected rollouts to the unchanged clipped GRPO objective (Shao et al., 2024; Schulman et al., 2017). The optimizer receives advantage $b_i(S_m) = a_i(S_m)$ in the ordinary condition and $b_i(S_m) = a_i(N)$ in the fixed-normalization control. For valid tokens, define $\rho_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x, y_{i,<t})/\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})$, where θ_{old} generated the current group. The optimizer minimizes

$$\mathcal{L}_{\text{GRPO}}(\theta; S, b) = -\frac{1}{|S|}\sum_{i\in S}\frac{1}{L_i}\sum_{t:m_{i,t}=1}\min\Big\{\rho_{i,t}(\theta)b_i(S),\ \text{clip}\big(\rho_{i,t}(\theta), 1-\eta, 1+\eta\big)b_i(S)\Big\} + \lambda_{\text{KL}}\,\mathcal{K}(\pi_\theta, \pi_{\text{ref}}). \tag{7}$$

Here $\eta > 0$ is the clipping radius and $\lambda_{\text{KL}} \geq 0$ weights the nonnegative KL regularizer $\mathcal{K}$ against reference policy π_{ref}. The leading minus sign converts the reward surrogate into a minimization loss. Gradient descent produces the policy update $\Delta\theta$, with no additional trainable selector module. Appendix A.4 specifies token masking and the KL estimate.

Selection procedure. Generate and verify the full group, construct q_i and z_i, and compute $\widetilde{g}_N$. If $\|\widetilde{g}_N\|_2 \leq \varepsilon_{\cos}$, set all utilities to zero and select the first m indices. Otherwise, compute attributions using the exact or sampled branch in equation 5 and select S_m. Record attribution and timing statistics, then minimize equation 7. Independent parameter-gradient diagnostics and downstream task performance assess the usefulness of this score-based selection.

3.8 UPDATE FIDELITY AND OPTIMIZATION PROGRESS

We now relate update fidelity to local progress on the full-group objective. Fix a rollout group sampled at θ_0 and its selected subset S, holding advantages and masks fixed. For rollout set T, let F_T be the frozen ascent surrogate, including its KL penalty, and $g_T = \nabla F_T(\theta_0)$ (Appendix D.1).

Theorem 1 (Progress from update fidelity)**.** *Suppose F_N has L-Lipschitz gradient on a convex neighborhood of θ_0, with $L > 0$, $G = \|g_N\| > 0$, and $\|g_S - \beta g_N\| \leq \delta < \beta G$ for some $\beta > 0$. If $\theta_+ = \theta_0 + g_S/(L\beta)$ lies in this neighborhood, then $F_N(\theta_+) - F_N(\theta_0) \geq (1-\epsilon^2)G^2/(2L)$, where $\epsilon = \delta/(\beta G)$.*

Proof sketch. Completing the square in the smoothness lower bound gives $\{G^2 - \|L\alpha g_S - g_N\|^2\}/(2L)$. Substituting $\alpha = 1/(L\beta)$ and the discrepancy bound proves the claim. Appendix D.4 gives the full proof and sharpness analysis.

The guarantee loses an ϵ^2 fraction of the full-gradient smoothness benchmark $G^2/(2L)$ for one ordinary gradient step on a frozen surrogate with an analytically calibrated step size. Appendix D.3 connects SSL's score proxy to the required gradient condition through a shared residual bound, and Appendix D.5 gives the ideal cosine interpretation. The construction in Appendix D.6 shows why this connection requires gradient calibration; Appendix D.7 gives the conditions for applying the bound.

4 EXPERIMENTS

We organize the experiments around five questions: **Q1.** Does selection improve task performance (Section 4.2)? **Q2.** Do the gains justify the selection cost (Section 4.3)? **Q3.** What information does selection preserve (Section 4.4)? **Q4.** Which rollout groups benefit most (Section 4.5)? **Q5.** How much valuation precision is useful (Section 4.6)?

4.1 EXPERIMENTAL SETUP

Benchmarks and metrics. We evaluate GSM8K, MATH500, AMC23, AIME24/25, Minerva, and English text-only OlympiadBench mathematics (Cobbe et al., 2021; Hendrycks et al., 2021; Lewkowycz et al., 2022; He et al., 2024), plus TACO and a fixed LiveCodeBench split (Li et al., 2023; Jain et al., 2024). Mathematics uses mean single-response correctness (Avg@R), with AIME pass@8 in resource comparisons. Code uses pass@1,5; the separate TACO-279 study uses mean

Table 1: **Transfer performance and cost at update 400.** Qwen2.5-3B-Instruct, GSM8K training, LoRA, $64 \rightarrow 16$, a 384-token response cap, and one L40S. Accuracy is Avg@5 (%); cost covers 400 updates. Additional conventions are given in Appendix B.5.

	Task performance ↑			Training cost ↓	
Method	AMC23	MATH500	Minerva	GPU-hours	Selection (%)
Random (Xu et al., 2025)	11.00	35.89	9.93	**4.5**	**0.01**
Max-Reward (Xu et al., 2025)	11.20	37.10	10.04	**4.5**	<u>0.03</u>
PODS (Xu et al., 2025)	11.50	38.27	10.29	4.8	0.12
D3S-Sample (Wang et al., 2026)	<u>12.50</u>	<u>38.90</u>	10.38	4.8	0.12
CPPO-Top-m (Lin et al., 2025)	11.30	38.55	10.31	<u>4.7</u>	0.06
CBS*-quota (Lu et al., 2026)	11.70	38.73	<u>10.41</u>	5.0	0.41
SSL (ours)	**12.80**$_{+0.30}$	**39.15**$_{+0.25}$	**10.60**$_{+0.19}$	5.2	1.92

test-case pass rate. We measure training GPU-hours, selector overhead, and independent gradient-sketch fidelity. Appendix B specifies cohorts, metrics, and hyperparameters.

Comparison protocol. The controlled GRPO comparison (Shao et al., 2024) retains $m = 16$ of $n = 64$ rollouts per prompt against Random, Max-Reward, PODS (Xu et al., 2025), and fixed-quota CPPO and CBS* adaptations (Lin et al., 2025; Lu et al., 2026). Native D3S, CBS/CBS*, and ARRoL (Wang et al., 2026; Lu et al., 2026; Xu et al., 2026) are compared under separately matched resource budgets. Comparisons share initialization, data, verification, decoding, and optimization settings; supplementary methods and optimizer controls appear in Appendix B.2.

Backbones and implementation. We use Qwen2.5-3B-Instruct, Qwen2.5-Math-7B, and Qwen3-4B-Base (Qwen et al., 2024; Yang et al., 2024; 2025). Short-response runs use LoRA on one L40S; extended runs train all parameters on eight A100-80GB GPUs. SSL uses $d = 128$, exact attribution for $n \leq 16$, and $K = 512$ permutations otherwise. Extended results report mean $\pm$ SD over five independent training seeds; short-response task summaries provide descriptive five-seed means without seed-level dispersion. Training costs include selection.

4.2 Q1: Does selection improve task performance?

Controlled mathematics selection. At the prespecified final checkpoint (update 400), SSL achieves the highest short-response means on AMC23, MATH500, and Minerva, exceeding PODS by 1.30, 0.88, and 0.31 percentage points, respectively (Table 1). Margins over each task's strongest alternative are 0.30, 0.25, and 0.19 points. The shared $64 \rightarrow 16$ quota isolates the effect of subset composition. Gains across all three benchmarks after GSM8K training also support transfer beyond the training task. Appendix C.2 reports both rollout budgets and complete trajectories.

Transfer across model backbones. SSL leads the seven-benchmark mean on all three backbones, exceeding the strongest alternative, CBS*-quota, by 0.49, 0.62, and 0.62 points, respectively (Figure 2; absolute means and SDs in Table 6). CBS*-quota remains higher on GSM8K for the two general-purpose backbones and on AMC23 for Qwen2.5-Math-7B, indicating task-dependent gains.

Code and optimizer transfer. SSL also leads TACO and LiveCodeBench pass@1,5, with LiveCodeBench pass@1 gains of 0.75/0.80 points over CBS*-quota on Qwen2.5-3B/Qwen3-4B. Under DAPO and GSPO, it improves the seven-benchmark mean over PODS by 0.75 points each (Tables 9–10). Appendix C.4 details code transfer; Q2 assesses the additional cost.

Takeaway Q1: Select for the joint update

A fixed rollout quota makes responses compete for update capacity; selecting them by individual reward can crowd out complementary contributions. The transfer results support valuing each response by what it adds to the group across backbones and policy objectives.

Table 2: **Native systems under separately matched resource budgets.** Qwen3-4B-Base, the extended math split, a 3,072-token response cap, and eight A100-80GB GPUs. Each budget block is a separate comparison. Mean denotes the seven-benchmark score shown in Figure 2. AIME denotes the average of AIME24/25 pass@8, estimated from 32 responses per problem. Scores are mean ± SD over five training seeds (%, ↑); bold and underline mark the best and second-best means per column. Budget accounting is detailed in Appendix B.5.

	Equal generated tokens		Equal GPU-hours		Equal update tokens	
Method	Mean	AIME	Mean	AIME	Mean	AIME
GRPO (Shao et al., 2024)	41.18±0.56	54.20±2.60	41.15±0.58	54.05±2.55	40.65±0.62	52.90±2.75
PODS (Xu et al., 2025)	41.82±0.60	56.90±2.35	41.80±0.60	56.75±2.40	41.88±0.59	56.95±2.30
D3S (Wang et al., 2026)	42.11±0.57	57.55±2.25	42.03±0.61	57.30±2.30	42.18±0.58	57.65±2.20
CBS* (Lu et al., 2026)	42.06±0.56	57.10±2.30	42.00±0.59	56.95±2.32	42.12±0.55	57.20±2.25
CBS (Lu et al., 2026)	42.34±0.54	58.00±2.18	42.26±0.56	57.80±2.22	42.40±0.53	58.10±2.15
ARRoL (Xu et al., 2026)	<u>42.41±0.55</u>	**59.35±2.10**	<u>42.38±0.57</u>	**59.10±2.15**	<u>42.47±0.54</u>	<u>59.05±2.08</u>
SSL (ours)	**42.72±0.52**	<u>59.20±2.12</u>	**42.56±0.55**	<u>58.95±2.18</u>	**42.80±0.51**	**59.35±2.05**

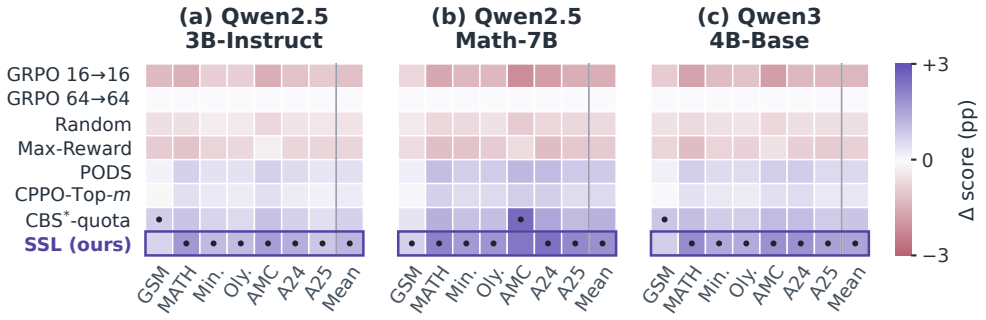

Figure 2: **Extended mathematical reasoning across three backbones.** Colors show mean-score changes (percentage points) from GRPO 64 → 64 for each backbone and metric at update 400, on a shared scale. Dots mark the best means; outlines identify SSL. The first four benchmarks use Avg@4 and the next three Avg@32; Mean averages all seven. Table 6 retains absolute means and SDs over five seeds; Appendix B.5 gives reporting conventions.

4.3 Q2: Do the gains justify the selection cost?

Cost at a fixed update horizon. At update 400, SSL uses 5.2 GPU-hours versus PODS's 4.8 (+8.3%), with 1.92% selection-only overhead (Table 1). Code and optimizer-transfer costs also increase; Appendix C.4 reports the breakdown.

Matched system budgets. Against native systems, SSL leads the seven-benchmark mean under separately matched generated-token, GPU-hour, and update-token budgets, exceeding ARRoL by 0.31, 0.18, and 0.33 points (Table 2). ARRoL leads AIME pass@8 by 0.15 points under the first two budgets; SSL leads under equal update tokens.

Geometry and allocation alternatives. Supplementary comparisons also favor SSL over SALT, GeoAlign, and VIP at matched GPU-hours, including auxiliary costs and native method choices (Table 10).

Takeaway Q2: Budget selection with training

Minimizing selection time can leave compute tied up in less informative updates. The matched-budget gains support allocating part of the training budget to better subset composition, with generation, selection, and updates costed together.

4.4 Q3: What information does selection preserve?

Geometry beyond reward scores. On 124 nonzero-reference Qwen3-4B-Base groups ($64 \to 16$, binary rewards), we measure reward extremity and directional novelty for signed features $s_i = a_i(N)z_i$:

$$E_i = |a_i(N)|, \qquad D_i = 1 - \max_{j \neq i} \frac{1 + \cos(s_i, s_j)}{2}. \tag{8}$$

Larger D_i means lower similarity to the nearest signed direction. Figure 3 compares selected-rollout densities; Appendix B.7 specifies estimation and aggregation.

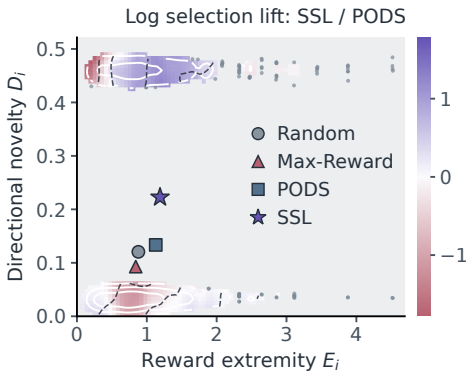

Figure 3: **Selection geometry under binary rewards.** Purple and rose denote positive and negative SSL-to-PODS log density ratios. White contours show all-rollout density; dashed contours show mean raw Shapley estimates. Gray masks low-support bins; dots show sparse observations. Markers denote selected-set centroids.

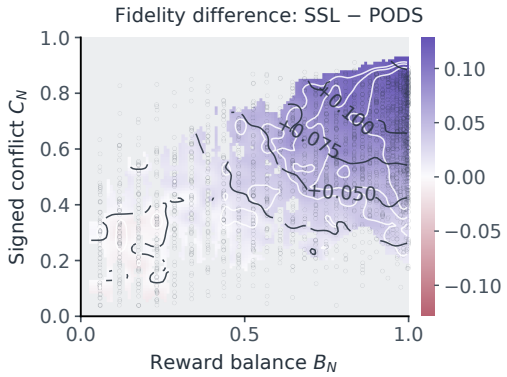

Figure 4: **Gradient-fidelity gain across prompt groups.** Color shows the smoothed SSL-minus-PODS cosine difference: purple favors SSL, and rose favors PODS. Points denote groups; white and dark contours show density and mean gain. Gray denotes insufficient support.

Compared with PODS, SSL favors the higher-novelty band, especially at intermediate reward extremity: mean novelty rises from 0.1335 to 0.2226 (+66.8%) while mean reward falls from 0.5181 to 0.4380 (Table 14).

Independent gradient and normalization controls. With selected indices fixed on 122 frozen Qwen3-4B groups, independent parameter-gradient sketches favor SSL over CBS*-quota under subset (cosine 0.79 versus 0.74) and full-group normalization (0.82 versus 0.77), with lower relative errors (Table 11). The gain persists with reward recentering fixed; short-response fidelity also improves at both rollout budgets (Appendix C.3).

Takeaway Q3: Preserve informative negative feedback

A zero-reward response can supply a distinctive negative-advantage direction that helps preserve the group update. Filtering solely for correctness can remove this useful contrast.

4.5 Q4: Which rollout groups benefit most?

Separating reward balance from directional conflict. We compare SSL and PODS on 2,483 eligible Qwen3-4B-Base groups across five training seeds at update 400 ($64 \to 16$). For mean binary reward $\bar{r}_N$ and signed contributions $s_i = a_i(N)z_i$, define

$$B_N = 4\bar{r}_N(1 - \bar{r}_N), \qquad C_N = 1 - \frac{\| \sum_{i \in N} s_i \|_2}{\sum_{i \in N} \|s_i\|_2 + \epsilon}. \tag{9}$$

Here B_N measures reward balance and C_N directional cancellation. Fidelity gain Δ_{fid} is the SSL-minus-PODS cosine difference against the full-group gradient, using independent sketches and subset-normalized advantages (Appendix B.8).

Table 3: **SSL permutation convergence on large groups.** $n = 64$, $m = 16$, $d = 128$. Rank and subset agreement use an independent $K = 8192$ Monte Carlo reference. RMS-SE summarizes attribution standard errors; cosine is external gradient fidelity. Values are mean ± SD over ten seed-level group averages; V_ρ counts valid group–seed pairs for rank correlation.

		Agreement with reference		Attribution	External		Selector
Method	K	Spearman ↑	Overlap ↑	RMS-SE ↓	Cosine ↑	V_ρ	ms/group ↓
SSL	32	$0.731_{\pm 0.021}$	$0.736_{\pm 0.015}$	$0.0476_{\pm 0.0032}$	$0.748_{\pm 0.007}$	1201	$16.7_{\pm 0.4}$
SSL	64	$0.811_{\pm 0.016}$	$0.801_{\pm 0.012}$	$0.0338_{\pm 0.0023}$	$0.761_{\pm 0.006}$	1210	$23.5_{\pm 0.5}$
SSL	128	$0.876_{\pm 0.011}$	$0.854_{\pm 0.010}$	$0.0240_{\pm 0.0016}$	$0.774_{\pm 0.005}$	1216	$36.8_{\pm 0.6}$
SSL	256	$0.925_{\pm 0.008}$	$0.899_{\pm 0.008}$	$0.0168_{\pm 0.0011}$	$0.784_{\pm 0.004}$	1219	$63.9_{\pm 0.8}$
SSL (ours)	**512**	$\mathbf{0.958_{\pm 0.005}}$	$\mathbf{0.929_{\pm 0.006}}$	$\mathbf{0.0119_{\pm 0.0008}}$	$\mathbf{0.790_{\pm 0.003}}$	**1220**	$\mathbf{118.4_{\pm 1.1}}$
SSL	1,024	$0.977_{\pm 0.003}$	$0.950_{\pm 0.004}$	$0.0085_{\pm 0.0006}$	$0.793_{\pm 0.003}$	1220	$227.6_{\pm 1.7}$
SSL	2,048	$0.988_{\pm 0.002}$	$0.966_{\pm 0.003}$	$0.0060_{\pm 0.0004}$	$0.796_{\pm 0.002}$	1220	$446.8_{\pm 2.9}$

Where selection changes gradient fidelity. SSL raises mean cosine from 0.7223 to 0.7844 ($\Delta_{\text{fid}} = 0.0621$). Of the four regions, the largest gain (0.1091) occurs when both balance and conflict are high ($B_N > 0.75$, $C_N > 0.65$; $n = 744$; Figure 4). Gains are smaller when only one coordinate is high and near zero when both are low (Table 15).

Takeaway Q4: Use conflict to guide selection effort

Reward-balanced groups can differ sharply in directional cancellation. This suggests allocating selection effort across prompts using both quantities, concentrating computation where balanced rewards and strong conflict coincide.

4.6 Q5: How much valuation precision is useful?

Representation size and sampling effort. On the frozen groups from Q3, increasing d from 32 to 128 raises independent gradient cosine from 0.72 to 0.79; $d = 256$ reaches 0.80 at 196.5 versus 118.4 ms per group (Table 12). At $d = 128$, increasing K from 512 to 2,048 raises cosine from 0.790 to 0.796 at 3.77× the selector cost (Table 3). These diminishing gains support the default $d = 128$, $K = 512$; Appendix C.6 gives convergence details.

Calibration against exact attribution. On 114 small groups ($n = 16$, nonconstant attributions), sampling RMSE relative to exact attributions falls from 0.0449 at $K = 32$ to 0.0056 at $K = 2{,}048$ (Table 13). At $K = 512$, RMS-SE (0.0114) closely matches RMSE (0.0112). This agreement suggests that estimated uncertainty can help gauge sampling precision. The calibration concerns the sampled approximation to the proxy game; SSL uses exact enumeration for these small groups during training.

Takeaway Q5: Stop when the selected direction stabilizes

Further attribution accuracy has diminishing value once the retained direction stabilizes. Sampling budgets can therefore be calibrated using selected-set stability together with independent gradient fidelity.

5 Conclusion

We studied how rollout down-sampling can retain the learning signal of a complete response group under a fixed policy-update budget. SSL addresses this problem by valuing responses through their contributions in different subsets, using compact score representations while leaving the policy objective unchanged. The experiments show improved mean reasoning performance across the tested models, while gradient diagnostics identify the largest fidelity gains in groups with balanced rewards and strong directional conflict. Together with the conditional analysis of local optimization progress, these findings motivate evaluating response selection by the learning directions it preserves, alongside reward statistics and computational cost. A promising direction is to investigate how contribution estimates can guide adaptive update budgets across prompts and stages of training.

AI USE STATEMENT

We used large language models (LLMs) to assist with literature search and synthesis, checking and refining mathematical formulas, organizing and presenting experimental results, and polishing the manuscript's language. The authors take full responsibility for the final content of this work, including all claims, results, and AI-assisted material.

REFERENCES

Powei Chang, Jinpeng Zhang, Chaoqun Sun, MiniWell Tsao, Lianrui Li, Jianxiang Xiang, Chenyu Wang, Yukang Gao, and Dongying Kong. SALT: When more rollouts don't help in group-based policy optimization and how to make them matter. *arXiv preprint arXiv:2606.05800*, 2026. URL https://arxiv.org/abs/2606.05800.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL https://arxiv.org/abs/2402.14008.

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. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

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. *arXiv preprint arXiv:2403.07974*, 2024. URL https://arxiv.org/abs/2403.07974.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474. PMLR, 2021. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.14858.

Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. TACO: Topics in algorithmic COde generation dataset. *arXiv preprint arXiv:2312.14852*, 2023. URL https://arxiv.org/abs/2312.14852.

Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. CPPO: Accelerating the training of group relative policy optimization-based reasoning models. *arXiv preprint arXiv:2503.22342*, 2025. URL https://arxiv.org/abs/2503.22342.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. URL https://arxiv.org/abs/2503.20783.

Xiaodong Lu, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, Zhijun Chen, Yu Luo, Fuzhen Zhuang, Yikun Ban, and Deqing Wang. Contextual rollout bandits for reinforcement learning with verifiable rewards. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.08499.

Hieu Trung Nguyen, Bao Nguyen, Wenao Ma, Yuzhi Zhao, Ruifeng She, and Viet Anh Nguyen. Adaptive rollout allocation for online reinforcement learning with verifiable rewards. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.01601.

Qwen et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Chao Wang, Tao Yang, Hongtao Tian, Yunsheng Shi, Qiyao Ma, Xiaotao Liu, Ting Yao, and Wenbo Ding. Learning more with less: A dynamic dual-level down-sampling framework for efficient policy optimization. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.22115.

Haobo Xu, Sirui Chen, Ruizhong Qiu, Yuchen Yan, Chen Luo, Monica Xiao Cheng, Jingrui He, and Hanghang Tong. Prune as you generate: Online rollout pruning for faster and better RLVR. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13876–13893. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.632. URL https://aclanthology.org/2026.acl-long.632/.

Y. E. Xu, Y. Savani, F. Fang, and J. Z. Kolter. Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning. In *arXiv preprint arXiv:2504.13818*, 2025.

An Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024. URL https://arxiv.org/abs/2409.12122.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Ningyuan Yang, Weihua Du, Weiwei Sun, Sean Welleck, and Yiming Yang. GradAlign: Gradient-aligned data selection for LLM reinforcement learning. *arXiv preprint arXiv:2602.21492*, 2026. URL https://arxiv.org/abs/2602.21492.

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

Wentao Zhang, Haoyu Zhang, Xinke Jiang, Yuxuan Cheng, Yuhan Pan, Miao Li, Zhipeng Qiao, Tao Feng, Zhen Tao, and Dengji Zhao. Fishing out free riders: Shapley-based reward attribution for parallel reasoning via reinforcement learning. *arXiv preprint arXiv:2607.18979*, 2026. URL https://arxiv.org/abs/2607.18979.

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

Ting Zhou, Zhenqing Ling, Yiyang Zhao, Ying Shen, and Daoyuan Chen. GeoAlign: Geometric rollout curation for robust LLM reinforcement learning. *arXiv preprint arXiv:2606.26917*, 2026. URL https://arxiv.org/abs/2606.26917.

Appendix Contents

A Technical Details 13

A.1 Token masks and score pooling . . . 13
A.2 Coalition evaluation and Monte Carlo bookkeeping . . . 13
A.3 Fixed-normalization control . . . 13
A.4 GRPO token objective . . . 13
A.5 Determinism and cost accounting . . . 14

B Experimental Details 14

B.1 Datasets and evaluation cohorts . . . 14
B.2 Comparators and fairness controls . . . 15
B.3 Metrics, uncertainty, and checkpoints . . . 16
B.4 Backbones, hyperparameters, and implementation . . . 16
B.5 Reporting conventions and comparison protocols . . . 18
B.6 Frozen-checkpoint diagnostic protocols . . . 18
B.7 Selection-landscape estimation and aggregation . . . 19
B.8 Reward-balance and conflict diagnostic protocol . . . 20

C Additional Experimental Results 21

C.1 Complete three-backbone mathematical results . . . 21
C.2 Complete checkpoint trajectories . . . 22
C.3 Gradient fidelity and runtime . . . 23
C.4 Extended code and paired optimizer comparisons . . . 23
C.5 Paired normalization and representation controls . . . 25
C.6 Permutation convergence and exact calibration . . . 25
C.7 Selected-set geometry and attribution . . . 26
C.8 Gradient fidelity by reward balance and conflict . . . 27

D Proofs 27

D.1 Frozen objective and the clipping neighborhood . . . 28
D.2 Exact coefficient geometry under subset normalization . . . 29
D.3 Shared residuals and joint update discrepancy . . . 30
D.4 Smoothness and proof of Theorem 1 . . . 31
D.5 Calibration, certificates, and the ideal cosine case . . . 32
D.6 Score and gradient fidelity in an explicit policy family . . . 34
D.7 Conditions for applying the guarantees . . . 35

A TECHNICAL DETAILS

A.1 TOKEN MASKS AND SCORE POOLING

For rollout i, the vector ℓ_i stores old-policy log-probabilities at a common maximum length T. The binary mask m_i is one for generated response tokens used by the policy objective and zero for padding, prompt tokens, and verifier-only positions. Thus $L_i = \mathbf{1}^\top m_i$ is the number of valid response tokens. The normalization in q_i makes the score magnitude comparable across response lengths before the fixed pooling map is applied.

The pooling matrix P is fixed before training. In the block-pooling implementation, the T token positions are partitioned into d contiguous bins, and row j of P assigns the same weight to positions in bin j and zero weight elsewhere. The row weights sum to one. The map is independent of rewards, validation results, and test statistics during selection. Varying d tests the effects of representation size on utility values, selected rollouts, and pooling cost.

A.2 COALITION EVALUATION AND MONTE CARLO BOOKKEEPING

Exact enumeration represents a coalition by an integer mask. For each mask, the implementation maintains the reward sum, squared-reward sum, score-vector sum, and reward-weighted score-vector sum. These aggregates recover μ_S, σ_S, and $Z^\top \mathbf{a}(S)$ without materializing an n-dimensional advantage vector for every coalition. The cache stores scalar utilities $v(S)$ and the aggregate quantities required for the next mask.

For a group with $n \le n_{\text{exact}}$, all 2^n utilities are computed once. Marginal differences are then obtained by pairing each mask that excludes i with the mask obtained by setting bit i. The scalar lookup contributes an $n2^n$ term, while score aggregation contributes a $2^n d$ term, giving the main-text bound $\mathcal{O}(2^n(d+n))$.

For a larger group, one recorded seed generates K permutations shared by all rollouts in that group. Each permutation yields a nested prefix sequence. The implementation evaluates the utility of each prefix once and adds the adjacent-prefix difference to the corresponding rollout. For rollout i, the Monte Carlo standard error is $\sqrt{s_i^2/K}$, where s_i^2 is the sample variance of its K marginal values. This uncertainty is logged together with the ranking and the selected-set overlap at the chosen budget.

A zero reference direction produces zero utility for every coalition. The resulting attributions are all zero, so the prescribed index tie-break selects the first m rollouts. A constant-reward coalition has zero centered advantages and receives utility zero. A coalition with a score vector below $\varepsilon_{\cos}$ also receives utility zero. These rules are shared by exact enumeration and permutation sampling.

A.3 FIXED-NORMALIZATION CONTROL

The selector always computes $\widetilde{\phi}_i$ from the ordinary coalition utility in equation 4. The control changes the advantages supplied to the downstream update after S_m has been selected. Specifically, the ordinary condition recomputes μ_{S_m} and σ_{S_m}, while the control uses the already computed μ_N and σ_N for every selected rollout. The rollout group, attribution values, permutation seed, selected indices, optimizer, and update budget remain paired.

This comparison isolates the effect of subset reward statistics at the update stage while holding the selection game fixed.

A.4 GRPO TOKEN OBJECTIVE

The ratio $\rho_{i,t}(\theta)$ in equation 7 uses the old-policy log-probability cached during rollout generation. The token sum is restricted to the mask m_i, and the factor L_i^{-1} gives each selected rollout equal weight independent of its response length. The KL term is estimated on the same valid response tokens against the frozen reference policy. The loss is minimized with the existing optimizer and learning-rate schedule; SSL changes the selected index set and the supplied advantage vector only.

The clipped surrogate is finite for finite advantages and probability ratios. Its sign is fixed by the leading negative sign in equation 7; the nonnegative KL term penalizes deviation from π_{ref}.

A.5 Determinism and cost accounting

Ties in $\widetilde{\phi}_i$ are resolved by increasing rollout index. Exact masks, permutation seeds, valid-token masks, and the values of $\varepsilon_{\mathrm{adv}}$, $\varepsilon_{\cos}$, n_{exact}, and K are recorded with each selection event. Selection time includes score pooling, coalition aggregation, utility evaluation, marginal accumulation, and logging. End-to-end training cost also includes rollout generation and the GRPO update; the frozen-group timing protocol is specified in Appendix B.6.

B Experimental Details

B.1 Datasets and evaluation cohorts

Short-response and extended studies. The short-response study trains Qwen2.5-3B-Instruct on GSM8K with a 384-token response cap for 400 updates, using five training seeds and $64 \to 16$ down-sampling. It evaluates AMC23, MATH500, and Minerva at updates 100, 200, 300, and 400 with five stochastic responses per problem. Additional $16 \to 8$ experiments include GSM8K evaluation and a 279-problem standard-I/O TACO subset. These experiments form the short-response cohort, with its own training corpus, response cap, decoding budget, and TACO subset.

Training data. For the extended mathematics experiments, we use the DAPO-Math-17k and MATH levels 3–5 training mixture used by CBS (Lu et al., 2026), starting from its released data manifests. We remove evaluation overlap and duplicate problems before reserving 512 training problems for validation; the training count is determined after filtering. The short-response cohort uses GSM8K training. Code models are trained separately on the official TACO training partition, with 512 eligible training problems reserved for validation. No evaluation solutions or tests are used to construct training rewards.

Benchmark coverage. Table 4 specifies the evaluation suite. We keep the TACO-279 subset identifier and manifest separate from the official full test partition. The extended TACO evaluation uses the official test partition with its supported execution modes. LiveCodeBench uses `release_v6`, restricted to problems dated 2024-08-01 through 2025-04-30, with a frozen problem-ID manifest. This temporal split fixes the evaluation cohort; overlap with each backbone's pretraining data is unverified. We exclude multimodal/physics OlympiadBench tasks because the present verifier and models target text-only mathematical reasoning.

Table 4: Evaluation cohorts and decoding budgets. Extended math uses temperature 1.0 and top-$p = 1.0$; extended code uses temperature 0.2 and top-$p = 1.0$. All methods within a row use the same parser, verifier, and token limits.

Benchmark	Role / split	Primary metric and samples
GSM8K	Official test; easy transfer control	Avg@4 in extended runs
MATH500	Standard 500-problem test subset	Avg@4; Avg@5 in short-response runs
AMC23	Competition mathematics	Avg@32; Avg@5 in short-response runs
AIME24 and AIME25	Separate competition test sets	Avg@32; mean pass@8 across both sets in Table 2
Minerva	272-problem mathematics set	Avg@4; Avg@5 in short-response runs
OlympiadBench	English, text-only mathematics	Avg@4
TACO	279-problem subset / official full test	Test-case pass rate / pass@1,5
LiveCodeBench	Fixed temporal slice of v6	pass@1,5 from 10 samples

Verification and leakage control. Mathematics receives a binary reward from one frozen answer-extraction and equivalence-checking pipeline, with failed parsing assigned zero. Code reward is the fraction of a fixed set of eight training tests passed. Training tests are selected deterministically per training problem. Code evaluation executes all available official tests, including special judges where supplied, in a sandbox with identical timeouts across methods. Compilation errors and timeouts count as failures. Dataset revisions, test-harness versions, split IDs, and prompt templates are recorded. Exact and normalized-text duplicate checks against every evaluation set precede training; unresolved near duplicates are logged for review.

B.2 Comparators and fairness controls

Fixed-pool selection. The main selection comparison holds n, m, the prompt batch, and the downstream GRPO loss fixed. Random samples m distinct indices uniformly. Max-Reward retains the highest rewards, while PODS maximizes the selected reward variance (Xu et al., 2025). CPPO-Top-m retains the largest absolute full-group advantages (Lin et al., 2025), disabling native threshold-based allocation. CBS*-quota ranks candidates using the online CBS scheduler but enforces exactly m responses per prompt and disables historical replay (Lu et al., 2026). CPPO-Top-m and CBS*-quota denote these fixed-quota adaptations. The released CBS* implementation can select across the current batch; its native version is therefore included separately. Deterministic score ties use rollout index; Random has its own recorded seed.

We also include GRPO-$16 \to 16$ as an equal-update-count anchor and GRPO-$64 \to 64$ as an equal-generation-count reference. The latter provides a full-group reference at higher update cost. Because response lengths vary, the equal-update-token panel counts actual non-padding tokens.

Interpreting D3S-Sample. D3S-Sample denotes the sample-selection condition of D3S. For one prompt, fixed full-group advantages $A_i = (r_i - \mu_N)/(\sigma_N + \varepsilon)$ satisfy

$$\mathrm{Var}_{i \in S}(A_i) = \frac{\mathrm{Var}_{i \in S}(r_i)}{(\sigma_N + \varepsilon)^2}. \tag{10}$$

Thus exact fixed-size, within-prompt variance maximization gives the same maximizing subsets as PODS, up to ties. Differences may arise from keeping full-group advantages, tie handling, or implementation details. Full D3S additionally changes selection across groups, token prioritization, and the training schedule (Wang et al., 2026); these components are preserved in its system comparison.

System comparisons. The primary recent-method comparison includes full D3S, native CBS/CBS*, and ARRoL. VIP provides a complementary adaptive-allocation comparison. We rerun these methods on the same checkpoint, training split, verifier, context limit, and hardware as their paired base optimizer. Native rollout allocation, replay, token masking, or early termination is preserved and explicitly recorded. ARRoL is evaluated without quality-head-assisted test-time voting in the primary accuracy table. Its discarded prefixes and head-training warm-up count toward cost. CBS replay storage and scheduler updates, and VIP's allocation warm-up, also count toward cost.

The system panel uses Qwen3-4B-Base and the extended math split. Separate panels compare performance against generated tokens, GPU-hours, and non-padding update tokens, ending at the largest budget reached by every method without extrapolation. Each panel matches one resource independently. Native allocation schedules are logged; SSL and PODS use $64 \to 16$.

Geometry and optimizer controls. SALT reweights updates, whereas GeoAlign uses learned geometric curation and can replace responses (Chang et al., 2026; Zhou et al., 2026). We compare them with their paired base optimizer in a supplementary panel, logging update multiplicities and all auxiliary-model costs. A Random/PODS/SSL selection comparison under DAPO and GSPO tests optimizer transfer (Yu et al., 2025; Zheng et al., 2025); each optimizer retains its own loss and clipping conventions. The Dr. GRPO versus Dr. GRPO+VIP pair follows VIP's allocation setting (Liu et al., 2025; Nguyen et al., 2026). Optimizer-specific settings are frozen using training validation data and applied identically within each pair. Each pair evaluates the combined effects of the method's selection, weighting, and allocation choices.

Normalization control. For every frozen diagnostic group, each selector first produces its standard indices S. We then evaluate these same indices with either $a_i(S)$ or $a_i(N)$ in the downstream

update. The coalition game and its Shapley values remain fixed. Backbone state, rollout tokens, gradient representation, and optimizer state are shared across each diagnostic pair. Independent training runs share initialization and prompt order; their rollout contents naturally diverge as their policies change.

B.3 METRICS, UNCERTAINTY, AND CHECKPOINTS

Task performance. For Q test problems, R independently sampled responses, and binary correctness y_{qr}, we report

$$\mathrm{Avg@}R = \frac{1}{QR}\sum_{q=1}^{Q}\sum_{r=1}^{R} y_{qr}, \qquad \widehat{\mathrm{pass@}k} = \frac{1}{Q}\sum_{q=1}^{Q}\left[1 - \frac{\binom{R-c_q}{k}}{\binom{R}{k}}\right], \tag{11}$$

where $c_q = \sum_r y_{qr}$, $k \le R$, and $\binom{a}{k} = 0$ when $a < k$. Avg@R averages correctness over individual responses. The short-response experiments average five decodes and report Avg@5. Extended mathematics uses Avg@4 for GSM8K, MATH500, Minerva, and OlympiadBench, and Avg@32 for AMC23 and AIME24/25. In Table 2, AIME is the arithmetic mean of AIME24 and AIME25 pass@8, each estimated from 32 responses per problem. All displayed task scores are percentages.

For TACO-279, mean test-case pass rate averages the fraction of tests passed per generated program, with equal problem weight. A full-solve indicator requires every test to pass; only this indicator is used in code pass@k. Extended TACO and LiveCodeBench use ten responses per problem and report pass@1 and pass@5. The mathematical Mean is the unweighted average of the declared seven benchmark scores, counting AIME24 and AIME25 separately. Figure 2 displays all seven benchmarks, with absolute scores in Table 6; the native-system and paired-optimizer tables use this aggregate alongside their displayed task metrics.

External gradient fidelity. On frozen groups, let $h_i = J\nabla_\theta \log \pi_\theta(o_i \mid x)/L_i$, where J is a fixed random projection independent of SSL's token-score pooling. We compare $G_S = |S|^{-1}\sum_{i\in S} b_i h_i$ with $G_N = |N|^{-1}\sum_{i\in N} a_i(N) h_i$, using cosine similarity and $\|G_S - G_N\|_2/\|G_N\|_2$. These metrics evaluate fidelity in projected parameter-gradient space. The extended diagnostic uses a 2,048-dimensional seeded signed projection and a fixed parameter-block list. Groups with zero reference norm are counted separately; a zero selected vector against a nonzero reference receives cosine zero and relative error one. Appendix B.6 specifies the paired diagnostics and their aggregation.

Cost and statistical reporting. Training GPU-hours include rollout generation, verification, token-score extraction, pooling, attribution, sorting, policy updates, and any auxiliary-model training. Evaluation is excluded from the training cost. Generated and non-padding update tokens define the corresponding resource-matched panels, with replay and discarded prefixes accounted for as specified in Appendix B.5. Selection (%) in the short-response table is its reported selection-only fraction of complete step time. The frozen-group diagnostics report selector latency separately from diagnostic gradient computation and end-to-end training cost.

Extended downstream results report mean $\pm$ SD over five independent training seeds. The short-response task summaries provide descriptive five-seed means; seed-level dispersion is unavailable. Training seeds are the units of replication, with decodes and frozen groups treated as repeated measurements. The fixed-update endpoint is update 400; the short-response study also reports updates 100, 200, and 300. Resource comparisons use the common budgets stated in Table 2. The primary endpoint is fixed before test evaluation.

B.4 BACKBONES, HYPERPARAMETERS, AND IMPLEMENTATION

Backbone coverage and context. Qwen2.5-3B-Instruct connects the short-response study to PODS; Qwen2.5-Math-7B overlaps with D3S/VIP; Qwen3-4B-Base overlaps with CBS/ARRoL. The comparison spans three checkpoints within the Qwen family. All three participate in the main mathematical selection comparison; code extensions use Qwen2.5-3B-Instruct and Qwen3-4B-Base. Each comparison uses one fixed checkpoint revision, tokenizer, and prompt template. Instruction checkpoints use their native chat template; base checkpoints use one fixed plain completion template shared by all methods.

The common math configuration uses at most 1,024 prompt tokens and 3,072 response tokens, within a 4,096-token total context. We use the released Qwen2.5-Math-7B context configuration with its original RoPE scaling. Overlength training prompts are removed before fixing the split. Evaluation prompts longer than 1,024 tokens are kept intact and receive the remaining context as their response budget; prompts exhausting the context count as failures. Code extensions use 4,096 prompt and 4,096 response tokens on their two supported backbones. Scores with different caps remain in separate tables.

Table 5: Training and implementation settings for the short-response and extended configurations. Unless otherwise specified, paired comparisons use identical hyperparameters.

Setting	Short-response cohort	Extended math / code
Training updates	400	400
Training seeds	5	5 (0–4)
Candidate / retained rollouts	64/16	64/16
Effective prompts per update	4	16
Response cap (train / eval)	384 / 384	3,072 / 3,072 (math)
Code response cap (train / eval)	384 / 384	4,096 / 4,096
Adaptation	LoRA, rank/alpha 64/64	Full parameters
Learning rate	5×10^{-6}	1×10^{-6}
Optimizer	AdamW	AdamW
Adam betas / epsilon	$(0.9, 0.999)$ / 10^{-8}	Same
Weight decay / gradient clip	0.1 / 1.0	0.1 / 1.0
Warm-up / LR schedule	10 updates / constant	10 updates / constant
PPO epochs per rollout batch	1	1
GRPO clip (lower / upper)	0.2 / 0.2	0.2 / 0.2
KL / entropy coefficient	0 / 0	0 / 0
Training temperature / top-p	1.0 / 1.0	1.0 / 1.0
Evaluation temperature	1.0	1.0 math; 0.2 code
Precision	BF16; FP32 reductions	Same
Hardware per run	1 L40S, 48GB	8 A100, 80GB each
SSL d / n_{exact} / K	128 / 16 / 512	Same
SSL ε_{adv} / $\varepsilon_{\cos}$	10^{-8} / 10^{-12}	Same

Execution details. Table 5 summarizes the training and implementation settings. The short-response setup adapts attention and MLP linear projections with LoRA rank 64, scaling 64, and zero dropout. The extended implementation uses `verl`, `vLLM`, BF16, activation checkpointing, and fully sharded data parallelism. We use Python 3.10, PyTorch 2.6, CUDA 12.4, Transformers 4.51.3, and vLLM 0.8.5, following the released CBS environment; the environment manifest records tested package versions and source commits. We use response microbatches of one per GPU and accumulate gradients to cover the entire effective retained batch before one optimizer step. Prompt batch size, rollouts per prompt, and response microbatch size specify separate levels of batching. Every hardware comparison uses the same device type and count, with L40S and A100 results reported separately.

SSL implementation and diagnostic scope. We mask non-response positions, divide each padded old-policy log-probability vector by its valid response length, and apply the fixed contiguous-bin linear pooling matrix from the method section. Pooling is fixed before training and independent of rewards and evaluation data. Coalition moments, vector reductions, and Shapley accumulation use FP32. The default is $d = 128$, exact enumeration for $n \leq 16$, and $K = 512$ independent uniform permutations otherwise. Empty, zero-direction, and constant-reward coalitions have zero utility. If the full-group proxy norm is at most 10^{-12}, every attribution is zero and the first m rollout indices are selected, consistent with the method's deterministic tie rule.

The normalization, representation, and attribution diagnostics use frozen groups from one frozen Qwen3-4B-Base checkpoint. We vary $d \in \{32, 64, 128, 256\}$ at $K = 512$, and $K \in \{32, 64, 128, 256, 512, 1024, 2048\}$ at $d = 128$, using the protocol in Appendix B.6. A separate $K = 8192$ estimate provides a same-game large-group reference, and exact small-group attribution provides ground truth for the proxy game. These diagnostics assess representation fidelity and attribution accuracy at the specified checkpoint and response cap. The balance–conflict analysis in

Appendix B.8 uses 512 validation prompts per seed across five checkpoints. Hyperparameters are frozen before extended test evaluation.

B.5 Reporting conventions and comparison protocols

Tables 1, 7, and 8 report the short-response experiments at each rollout budget and checkpoint. Figure 2 and Tables 6, 2, 9, and 10 report the extended experiments, with accuracy summarized as mean $\pm$ SD over five training seeds. The short-response task summaries provide descriptive means; seed-level dispersion is unavailable. Bold indicates the best mean under each table's stated comparison rule, and shading identifies SSL regardless of rank; underlining marks the second-best means in Tables 6 and 2.

Table 1 conventions. Bold and underlining denote the best and second-best distinct values in each column, respectively, including ties; accuracy is maximized and cost is minimized. The pale purple row identifies SSL. Green lower-right annotations report its accuracy gains in percentage points over the strongest baseline in each benchmark: +0.30 over D3S-Sample on AMC23, +0.25 over D3S-Sample on MATH500, and +0.19 over CBS*-quota on Minerva. Selection is the reported selection-only fraction of step time; feature construction and other costs are also included in GPU-hours. Random and Max-Reward follow the down-sampling baselines of Xu et al. (2025). D3S-Sample is the sample-only condition associated with Wang et al. (2026); full D3S is evaluated in the system comparisons. CPPO-Top-m and CBS*-quota are the controlled adaptations defined in Appendix B.2, based on Lin et al. (2025) and Lu et al. (2026).

All comparisons use matched experimental runs. In particular, PODS reports a different length/reward configuration (Xu et al., 2025); D3S trains on DeepScaleR and evaluates 32 responses (Wang et al., 2026); CBS uses a 4,096-token response cap, eight rollouts per prompt, and evaluation-based checkpoint selection (Lu et al., 2026); ARRoL uses an 8,192-token sequence limit and pass@16 on AMC/AIME (Xu et al., 2026). CPPO, VIP, SALT, and GeoAlign also differ in the backbone, optimization, or decoding protocol (Lin et al., 2025; Nguyen et al., 2026; Chang et al., 2026; Zhou et al., 2026). These protocol differences motivate rerunning each method under shared conditions.

Figure 2 and Table 6 conventions. Olymp. denotes English text-only mathematics in OlympiadBench. All metrics estimate single-response correctness; Avg@32 averages correctness over 32 responses. GRPO follows Shao et al. (2024); Random, Max-Reward, and PODS follow Xu et al. (2025); adapted CPPO and CBS* are based on Lin et al. (2025); Lu et al. (2026). Mean includes all seven benchmarks. Figure 2 uses GSM, MATH, Min., Oly., AMC, A24, and A25 for GSM8K, MATH500, Minerva, OlympiadBench, AMC23, AIME24, and AIME25, respectively. Colors encode each reported mean minus its corresponding GRPO $64 \rightarrow 64$ mean on a shared percentage-point scale. Dots mark the best mean within each column, including ties; outlines identify SSL regardless of rank. The heatmaps encode means only. Table 6 reports the absolute scores and all reported SDs, including the SD of Mean.

Table 2 budgets and accounting. Matched budgets are 3.84×10^8 generated tokens, 320 A100-GPU-hours, and 9.60×10^7 non-padding update tokens, respectively. GRPO uses $64 \rightarrow 64$; PODS and SSL use $64 \rightarrow 16$. D3S, CBS/CBS*, and ARRoL retain native allocation, masking, replay, or pruning. Generated tokens include pruned prefixes, update tokens include replay/multiplicity, and GPU-hours include auxiliary training. ARRoL uses ordinary decoding without quality-head voting.

B.6 Frozen-checkpoint diagnostic protocols

These frozen-checkpoint diagnostics assess external gradient fidelity, attribution accuracy, and the conditions under which selection helps, without additional policy training. In Tables 3, 12, and 13, bold shaded rows identify the default configuration. The gray row in Table 13 is the exact reference; an em dash denotes an unavailable or inapplicable quantity.

Shared frozen groups. We use Qwen3-4B-Base GRPO-$64 \rightarrow 64$, training seed 0, update 400. We fix 128 validation prompts and generate 64 responses each with temperature/top-$p = 1.0$ and

a 3,072-token cap, caching tokens, rewards, old-policy log-probabilities, and independent 2,048-dimensional gradient sketches. The parameter blocks used for gradient sketching and the projection seed are fixed. Random and SSL use ten selection seeds (0–9); deterministic selectors retain their tie rule. These seeds repeat selection on the same frozen groups.

Normalization control. Table 11 evaluates each fixed S with $a_i(S)$ or $a_i(N)$, sharing G_N, $|S|^{-1}$ scaling, the coalition game, and the selected indices. We average the ten selections within each group, then report mean $\pm$ SD across the 122 groups with $\|G_N\|_2 > 10^{-12}$; six reference-zero groups are excluded. A selected-zero vector against a nonzero reference retains cosine zero and relative error one. The comparison measures fidelity in projected parameter-gradient space.

Representation dimension. Table 12 varies d at $K = 512$, using subset-normalized advantages and the same gradient eligibility mask and aggregation as Table 11. Overlap is $|S_d \cap S_{128}|/16$ with paired permutation seeds. Changing d changes the game, so overlap measures selected-set agreement with the default representation. The default row has unit overlap by definition.

Timing. We measure selector milliseconds per group on one A100-80GB with device-resident inputs, warm-up, synchronized timing, and identical caching. Timing includes masking, normalization, pooling, coalition computation, attribution, and selection, but excludes generation, verification, and diagnostic backward passes. Generation and verification remain included in the main tables' training GPU-hours.

Large-group convergence. Table 3 uses $n = 64$, $m = 16$, $d = 128$ and nested permutation-stream prefixes across K within each seed. A disjoint seed supplies a fixed approximate $K = 8192$ reference per group. Spearman uses average ranks for ties; overlap is $|S_K \cap S_{8192}|/16$. Both exclude the six groups with constant reference attributions, leaving 122 eligible groups and at most $122 \times 10 = 1220$ valid group–seed pairs. Constant estimates are additionally excluded from Spearman but retain their deterministic subsets for overlap. V_ρ counts valid rank-correlation pairs; values below 1220 reflect these additional exclusions. External cosine uses the nonzero-gradient mask and subset normalization from Table 11.

Monte Carlo uncertainty. For each group and selection seed, we summarize attribution standard errors by

$$\text{RMS-SE} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{s_i^2}{K} \right)^{1/2}, \qquad \text{RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} (\widetilde{\phi}_i - \phi_i)^2 \right)^{1/2}, \tag{12}$$

where s_i^2 is the marginal-contribution sample variance. RMSE requires exact ϕ_i and is used only for small groups. Tables 3 and 13 report mean $\pm$ SD of ten seed-level group averages; V_ρ is an integer. RMS-SE summarizes uncertainty in individual attribution estimates.

Exact small-group calibration. We take the first 16 indices of each cached group, independently of reward, with $m = 4$, $d = 128$, and enumerate all 2^{16} utilities once for exact Shapley values. For this diagnostic only, permutation sampling replaces the production exact branch at $n = 16$. Table 13 uses the 114 groups with nonconstant exact vectors for RMSE, RMS-SE, and overlap; Spearman also excludes constant estimates. Fourteen groups have constant exact vectors. The exact row has zero error and unit agreement by definition; latency is unavailable for this reference.

The corresponding control and attribution results appear in Appendices C.5 and C.6.

B.7 Selection-landscape estimation and aggregation

Records and eligibility. The analysis in Section 4.4 uses cached Qwen3-4B-Base GRPO-$64 \to 64$ rollouts at update 400 (training seed 0): 128 validation prompts with 64 responses each. Binary rewards are centered and scaled by the empirical population standard deviation plus $\varepsilon_{\text{adv}} = 10^{-8}$. Four groups have $\|\widetilde{g}_N\|_2 \leq 10^{-12}$ and zero attribution for every rollout; excluding them leaves 124 groups and 7,936 records. This score-space eligibility criterion differs from the independent gradient-sketch criterion in Appendix B.6.

Selections and aggregation. Each method records one subset of 16 rollouts per group, yielding $124 \times 16 = 1,984$ selected observations per method after filtering. Selected-set means pool these observations with equal weight; the equal group sizes and budgets also give equal group weight. Max-Reward uses decreasing reward with index-based ties. PODS maximizes binary reward variance under the fixed budget, using index-based ties within each reward class. SSL selects the largest raw Shapley estimates from $K = 512$ permutations. The landscape summarizes one recorded selection per method; the paired diagnostics average ten selection seeds.

Selection lift. For a bin B in the extremity–novelty plane, define

$$\Lambda(B) = \log \frac{p_{\mathrm{SSL}}(B) + \epsilon}{p_{\mathrm{PODS}}(B) + \epsilon}, \qquad p_M(B) = \Pr\big((E_i, D_i) \in B \mid i \in S_M\big), \tag{13}$$

with $\epsilon = 2 \times 10^{-5}$. Positive lift indicates relative enrichment in SSL's selected distribution compared with PODS.

Density estimation. The density visualization uses display coordinates, which differ from the unshifted reward extremity and novelty by less than 0.0043 and 0.000463, respectively. All reported means use unshifted values. We divide the displayed range into 84×72 bins; the upper limits are 1.04 times the maximum display extremity and $\min(1, 1.08 \max_i D_i^{\mathrm{disp}})$ for novelty. All-rollout counts and method-specific selected counts are smoothed with a Gaussian kernel of standard deviation 1.35 bins and reflecting boundaries. Each selected histogram is normalized to unit mass before applying Eq. (13), with $\epsilon = 2 \times 10^{-5}$. Bins with smoothed all-rollout count below one are masked, while sparse observations remain visible as gray dots. White contours use all-rollout density thresholds enclosing 50%, 80%, and 95% of the smoothed mass, restricted to supported bins. Centroids average the corresponding display coordinates across selected rollouts.

Raw attribution overlay. The overlay uses the raw signed Shapley estimates, without clipping or min–max normalization. In each bin, the smoothed attribution sum is divided by the smoothed all-rollout count. Contour levels include the 15th, 40th, 70th, and 90th percentiles of supported bin means, together with zero and a representative negative level when both signs occur.

Selected-set statistics and their interpretation are reported in Appendix C.7.

B.8 Reward-balance and conflict diagnostic protocol

Cohort and eligibility. Section 4.5 uses five extended-math Qwen3-4B-Base checkpoints at update 400 (training seeds 0–4), each with 512 held-out prompts and 64 rollouts per prompt. Both selectors retain 16 rollouts under the same downstream update rule. This 2,560-group cohort is distinct from the 128-prompt diagnostics in Appendices B.6 and B.7. Eligibility excludes constant rewards or a zero gradient reference. All 77 excluded groups have constant rewards and undefined conflict and cosine fields. The retained counts for seeds 0–4 are 497, 493, 497, 496, and 500.

Coordinates. With $\bar{r}_N = |N|^{-1} \sum_i r_i$ and binary rewards, B_N is zero for constant rewards and one for an equal split. Its symmetry between mostly correct and mostly incorrect groups makes it a measure of reward balance. The triangle inequality gives $C_N \in [0, 1]$, with $\epsilon > 0$ stabilizing the denominator. Larger values indicate stronger relative cancellation. Conflict depends on rewards through the advantages in s_i. Subset fidelity also depends on how selection and advantage renormalization change the signed aggregate.

External response. Let h_i denote the independent parameter-gradient sketch of rollout i. Using the full-group and subset advantages defined in the method, form

$$G_N = \frac{1}{|N|} \sum_{i \in N} a_i(N) h_i, \qquad G_S = \frac{1}{|S|} \sum_{i \in S} a_i(S) h_i. \tag{14}$$

The group-level response is

$$\Delta_{\mathrm{fid}}(N) = \cos(G_{S_{\mathrm{SSL}}}, G_N) - \cos(G_{S_{\mathrm{PODS}}}, G_N). \tag{15}$$

Both selectors share the reference; a zero selected vector has cosine zero against a nonzero reference. The response evaluates alignment in an independent parameter-gradient projection.

Surface estimation. The plot uses display coordinates within 0.001 of the original balance and conflict values; regional membership and reported means use the original values. We bin groups on a 110×100 grid over $[0, 1]^2$ and smooth counts and gain sums with the same Gaussian kernel (standard deviation 2.15 bins, truncated at four standard deviations, with zero padding). Their ratio estimates the local mean gain. Color is shown only where the smoothed count is at least 0.06 and the effective local count $(\sum_i w_i)^2 / \sum_i w_i^2$ is at least 10, where w_i is the kernel weight of group i. Unsupported cells are gray, with group locations still visible. White contours use the 55th, 75th, and 90th percentiles of supported smoothed counts; dark contours trace mean gain. The symmetric rose–white–indigo scale is centered at zero.

Regional summaries and per-seed gains are reported in Appendix C.8.

C Additional Experimental Results

C.1 Complete three-backbone mathematical results

Table 6 supplies the complete numerical results underlying Figure 2. The GRPO anchors match retained-response count ($16 \to 16$) or generated-response count ($64 \to 64$); all selectors use $64 \to 16$.

SSL achieves seven-benchmark means of 32.88, 50.33, and 42.88 for Qwen2.5-3B-Instruct, Qwen2.5-Math-7B, and Qwen3-4B-Base, respectively. Its lead over CBS*-quota is 0.49, 0.62, and 0.62 percentage points; the task-level exceptions are discussed in Section 4.2.

Table 6: **Complete numerical results for Figure 2.** Absolute scores are mean ± SD over five seeds (%); Mean is the unweighted mean of the seven benchmark scores. All runs use the same filtered training mixture, full-parameter adaptation, a 3,072-token response cap, and update 400. Bold and underline mark the best and second-best means per column within each backbone. Additional conventions are given in Appendix B.5.

		Avg@4 ↑				Avg@32 ↑			
Backbone	Method	GSM8K	MATH500	Minerva	Olymp.	AMC23	AIME24	AIME25	Mean
Qwen2.5 3B-Instruct	GRPO $16 \to 16$	$78.90_{\pm 0.47}$	$47.10_{\pm 0.75}$	$21.25_{\pm 0.87}$	$16.80_{\pm 0.90}$	$29.30_{\pm 1.40}$	$11.85_{\pm 1.43}$	$8.21_{\pm 1.11}$	$30.49_{\pm 0.83}$
	GRPO $64 \to 64$	$80.15_{\pm 0.49}$	$48.55_{\pm 0.80}$	$22.10_{\pm 0.79}$	$17.65_{\pm 0.76}$	$30.80_{\pm 1.36}$	$12.96_{\pm 1.24}$	$9.15_{\pm 0.96}$	$31.62_{\pm 0.61}$
	Random	$79.60_{\pm 0.57}$	$48.00_{\pm 0.79}$	$21.80_{\pm 1.12}$	$17.30_{\pm 0.90}$	$30.09_{\pm 1.23}$	$12.50_{\pm 1.40}$	$8.75_{\pm 1.25}$	$31.15_{\pm 0.69}$
	Max-Reward	$79.25_{\pm 0.60}$	$47.45_{\pm 0.73}$	$21.34_{\pm 1.09}$	$16.95_{\pm 0.98}$	$30.55_{\pm 1.60}$	$12.21_{\pm 1.57}$	$8.40_{\pm 1.41}$	$30.88_{\pm 0.74}$
	PODS	$80.30_{\pm 0.44}$	$49.30_{\pm 0.72}$	$22.56_{\pm 0.92}$	$18.10_{\pm 0.88}$	$31.59_{\pm 1.36}$	$13.54_{\pm 1.30}$	$9.54_{\pm 1.06}$	$32.13_{\pm 0.65}$
	CPPO-Top-m	$80.05_{\pm 0.44}$	$49.05_{\pm 0.71}$	$22.41_{\pm 0.94}$	$17.95_{\pm 0.73}$	$31.30_{\pm 1.12}$	$13.25_{\pm 1.12}$	$9.35_{\pm 1.24}$	$31.91_{\pm 0.29}$
	CBS*-quota	$\mathbf{80.95_{\pm 0.37}}$	$\underline{49.55_{\pm 0.72}}$	$\underline{22.79_{\pm 0.72}}$	$\underline{18.25_{\pm 0.82}}$	$\underline{31.84_{\pm 1.25}}$	$\underline{13.71_{\pm 1.25}}$	$\underline{9.65_{\pm 1.07}}$	$\underline{32.39_{\pm 0.44}}$
	SSL (ours)	$\underline{80.85_{\pm 0.41}}$	$\mathbf{50.35_{\pm 0.57}}$	$\mathbf{23.25_{\pm 0.77}}$	$\mathbf{18.80_{\pm 0.83}}$	$\mathbf{32.45_{\pm 1.28}}$	$\mathbf{14.35_{\pm 0.93}}$	$\mathbf{10.10_{\pm 0.90}}$	$\mathbf{32.88_{\pm 0.61}}$
Qwen2.5 Math-7B	GRPO $16 \to 16$	$90.60_{\pm 0.46}$	$68.50_{\pm 0.77}$	$36.19_{\pm 0.74}$	$30.70_{\pm 0.81}$	$55.09_{\pm 1.35}$	$27.79_{\pm 1.02}$	$19.60_{\pm 1.12}$	$46.92_{\pm 0.46}$
	GRPO $64 \to 64$	$91.30_{\pm 0.37}$	$70.10_{\pm 0.58}$	$37.50_{\pm 0.86}$	$32.00_{\pm 0.70}$	$57.20_{\pm 1.22}$	$29.60_{\pm 1.12}$	$21.10_{\pm 0.87}$	$48.40_{\pm 0.47}$
	Random	$90.95_{\pm 0.40}$	$69.40_{\pm 0.79}$	$36.86_{\pm 0.86}$	$31.50_{\pm 0.86}$	$56.30_{\pm 1.45}$	$28.90_{\pm 1.11}$	$20.40_{\pm 1.15}$	$47.76_{\pm 0.62}$
	Max-Reward	$90.70_{\pm 0.48}$	$68.95_{\pm 0.70}$	$36.40_{\pm 0.85}$	$31.10_{\pm 0.96}$	$56.59_{\pm 1.42}$	$28.40_{\pm 1.45}$	$20.10_{\pm 1.09}$	$47.46_{\pm 0.76}$
	PODS	$91.55_{\pm 0.43}$	$71.20_{\pm 0.56}$	$38.31_{\pm 0.69}$	$32.85_{\pm 0.81}$	$58.41_{\pm 1.24}$	$30.75_{\pm 1.02}$	$22.00_{\pm 0.82}$	$49.29_{\pm 0.57}$
	CPPO-Top-m	$91.40_{\pm 0.44}$	$70.85_{\pm 0.65}$	$38.05_{\pm 0.85}$	$32.60_{\pm 0.81}$	$58.00_{\pm 1.24}$	$30.40_{\pm 1.06}$	$21.65_{\pm 0.86}$	$48.99_{\pm 0.42}$
	CBS*-quota	$\underline{91.70_{\pm 0.37}}$	$\underline{71.45_{\pm 0.52}}$	$\underline{38.55_{\pm 0.70}}$	$\underline{33.10_{\pm 0.64}}$	$\mathbf{59.80_{\pm 1.13}}$	$\underline{31.10_{\pm 1.00}}$	$\underline{22.25_{\pm 0.99}}$	$\underline{49.71_{\pm 0.24}}$
	SSL (ours)	$\mathbf{92.05_{\pm 0.36}}$	$\mathbf{72.30_{\pm 0.51}}$	$\mathbf{39.25_{\pm 0.70}}$	$\mathbf{33.85_{\pm 0.58}}$	$\underline{59.59_{\pm 0.90}}$	$\mathbf{32.10_{\pm 1.07}}$	$\mathbf{23.15_{\pm 0.82}}$	$\mathbf{50.33_{\pm 0.47}}$

Continued on next page

Table 6: Complete numerical results for Figure 2 (continued).

Backbone	Method	Avg@4 ↑ GSM8K	Avg@4 ↑ MATH500	Avg@4 ↑ Minerva	Avg@4 ↑ Olymp.	Avg@32 ↑ AMC23	Avg@32 ↑ AIME24	Avg@32 ↑ AIME25	Mean
Qwen3 4B-Base	GRPO $16\to16$	$87.80_{\pm0.43}$	$60.90_{\pm0.74}$	$28.49_{\pm0.86}$	$24.80_{\pm0.99}$	$44.70_{\pm1.33}$	$19.79_{\pm1.30}$	$13.21_{\pm1.16}$	$39.96_{\pm0.70}$
	GRPO $64\to64$	$88.70_{\pm0.39}$	$62.60_{\pm0.79}$	$29.71_{\pm0.84}$	$25.90_{\pm0.70}$	$46.50_{\pm1.19}$	$21.10_{\pm1.18}$	$14.50_{\pm1.03}$	$41.29_{\pm0.59}$
	Random	$88.20_{\pm0.49}$	$61.95_{\pm0.83}$	$29.19_{\pm0.87}$	$25.45_{\pm0.98}$	$45.80_{\pm1.49}$	$20.54_{\pm1.32}$	$13.96_{\pm1.16}$	$40.73_{\pm0.63}$
	Max-Reward	$87.95_{\pm0.51}$	$61.40_{\pm0.80}$	$28.95_{\pm0.88}$	$25.10_{\pm0.84}$	$46.00_{\pm1.60}$	$20.25_{\pm1.36}$	$13.60_{\pm1.33}$	$40.46_{\pm0.63}$
	PODS	$88.90_{\pm0.47}$	$63.40_{\pm0.78}$	$30.29_{\pm0.75}$	$26.45_{\pm0.78}$	$47.30_{\pm1.33}$	$21.96_{\pm1.17}$	$15.10_{\pm0.97}$	$41.91_{\pm0.61}$
	CPPO-Top-m	$88.75_{\pm0.48}$	$63.05_{\pm0.80}$	$30.06_{\pm0.92}$	$26.20_{\pm0.92}$	$47.00_{\pm1.38}$	$21.65_{\pm1.21}$	$14.85_{\pm1.22}$	$41.65_{\pm0.42}$
	CBS*-quota	$\mathbf{89.65_{\pm0.43}}$	$63.70_{\pm0.69}$	$30.55_{\pm0.84}$	$26.70_{\pm0.77}$	$47.66_{\pm1.07}$	$22.21_{\pm0.98}$	$15.35_{\pm1.14}$	$42.26_{\pm0.59}$
	SSL (ours)	$89.50_{\pm0.44}$	$\mathbf{64.60_{\pm0.72}}$	$\mathbf{31.19_{\pm0.80}}$	$\mathbf{27.35_{\pm0.63}}$	$\mathbf{48.41_{\pm1.29}}$	$\mathbf{23.00_{\pm1.17}}$	$\mathbf{16.10_{\pm1.04}}$	$\mathbf{42.88_{\pm0.56}}$

C.2 Complete checkpoint trajectories

Table 7 reports all recorded checkpoints, with update 400 as the primary endpoint. TACO-279 reaches 27.61 at update 300 and 27.14 at the final checkpoint. The table reports five-seed task means; per-seed scores are unavailable for estimating SDs or paired confidence intervals.

At update 400, SSL exceeds PODS by 1.30, 0.88, and 0.31 percentage points on AMC23, MATH500, and Minerva; the corresponding gains over Random are 1.80, 3.26, and 0.67 points. On TACO-279, mean test-case pass rate rises from PODS's 25.76% to 27.14%, using the partial-credit metric defined in Appendix B.3.

Table 7: **Complete short-response task trajectories.** Qwen2.5-3B-Instruct with a 384-token response cap. Mathematical scores are Avg@5 (%); TACO-279 uses mean test-case pass rate (%), evaluated on all available tests. Values are the reported five-run means. Each block compares methods at the same rollout budget and task; columns are update counts. Update 400 is the primary endpoint.

Method	100	200	300	400
$64\to16$ *AMC23: Avg@5*				
Random	7.00	14.50	11.50	11.00
PODS	8.00	12.00	9.50	11.50
D3S-Sample	10.20	13.50	13.90	12.50
SSL (ours)	10.00	12.50	14.90	12.80
$64\to16$ *MATH500: Avg@5*				
Random	34.96	39.35	38.19	35.89
PODS	36.01	37.82	40.08	38.27
D3S-Sample	34.21	39.83	40.18	38.90
SSL (ours)	33.59	38.23	40.83	39.15
$64\to16$ *Minerva: Avg@5*				
Random	9.12	10.07	10.37	9.93
PODS	10.15	10.37	11.25	10.29
D3S-Sample	9.62	10.57	11.01	10.38
SSL (ours)	9.49	10.26	11.41	10.60
$16\to8$ *GSM8K: Avg@5*				
PODS	33.01	31.25	32.84	31.58
SSL (ours)	31.30	30.95	30.46	31.45
$16\to8$ *AMC23: Avg@5*				
PODS	12.50	10.00	10.00	10.50
SSL (ours)	11.50	14.00	12.50	10.00

Continued on next page

Table 7: Complete short-response task trajectories (continued).

Method	100	200	300	400
16→8 *MATH500: Avg@5*				
PODS	28.51	27.98	28.39	27.02
SSL (ours)	26.77	28.91	26.69	29.60
16→8 *TACO-279: test-case pass rate*				
PODS	25.16	24.15	24.60	25.76
SSL (ours)	24.59	26.57	27.61	27.14

C.3 Gradient fidelity and runtime

Table 8 reports gradient-sketch fidelity at each short-response rollout budget, with mean and SD over five runs. It measures alignment with the full-group gradient sketch. The extended experiments use a separate 2,048-dimensional projection, as specified in Appendix B.3.

Table 8: **Short-response gradient-sketch fidelity.** Reported mean ± SD over five runs on Qwen2.5-3B-Instruct. The full-group reference contains 16 or 64 rollouts, respectively. Bold denotes the best reported mean within each budget and metric.

	16→8		64→16	
Method	Cosine ↑	Rel. error ↓	Cosine ↑	Rel. error ↓
Random	0.62 ± 0.18	0.91 ± 0.25	0.64 ± 0.19	0.89 ± 0.24
PODS	0.69 ± 0.16	0.78 ± 0.22	0.71 ± 0.18	0.77 ± 0.21
D3S-Sample	0.68 ± 0.18	0.82 ± 0.26	0.69 ± 0.20	0.83 ± 0.23
SSL (ours)	$\mathbf{0.74} \pm 0.14$	$\mathbf{0.70} \pm 0.20$	$\mathbf{0.76} \pm 0.16$	$\mathbf{0.68} \pm 0.17$

For the $64 \to 16$ runtime study, the reported step times are 40.2, 43.3, 43.4, and 46.8 seconds for Random, PODS, D3S-Sample, and SSL, respectively. The SSL–PODS step-time increase is about 8.1%; the increase in their rounded total GPU-hours is about 8.3%. The SSL total of 5.2 GPU-hours is consistent with 400 updates at 46.8 seconds per step. Task accuracy at each checkpoint is reported in Table 7.

C.4 Extended code and paired optimizer comparisons

Table 9 evaluates the official TACO test split and the fixed LiveCodeBench temporal split. TACO-279 belongs to the separate short-response cohort. Code pass@k requires complete test-suite success, unlike the partial-credit metric in Table 7. Relative to PODS, SSL improves LiveCodeBench pass@1 by 0.95 and 1.10 points for Qwen2.5-3B-Instruct and Qwen3-4B-Base, with training costs increasing from 304.5 to 318.9 and from 321.2 to 336.1 GPU-hours, respectively.

Table 10 separates optimizer transfer, geometry-based methods, and adaptive allocation. Initialization, training data, response caps, evaluation, and hardware are shared within each block. At 320 A100-GPU-hours, the GRPO comparison gives means of 42.56 for SSL, 42.16 for GeoAlign, and 41.98 for SALT. At 312 A100-GPU-hours, the Dr. GRPO comparison gives 50.16 for SSL and 49.59 for VIP. Each budget includes auxiliary costs and native selection, weighting, or allocation choices. Under DAPO and GSPO, SSL improves the mean over PODS by 0.75 points each, using 338.5/346.0 versus 329.2/337.1 GPU-hours.

Table 9: **Extended code generation.** Separate models are trained on TACO and evaluated at update 400 with a 4,096-token response cap, temperature 0.2, and ten sampled programs per problem. TACO is the official test partition. LiveCodeBench is `release_v6`, dated 2024-08-01–2025-04-30. Scores are full-solve percentages (↑), reported as mean ± SD over five training seeds. All non-anchor selectors use $64 \rightarrow 16$; bold denotes the highest accuracy mean within each backbone.

	TACO		LiveCodeBench		Training
Method	pass@1	pass@5	pass@1	pass@5	GPU-hours ↓
Qwen2.5-3B-Instruct					
GRPO $16 \rightarrow 16$	$18.65_{\pm 0.74}$	$31.40_{\pm 1.10}$	$13.80_{\pm 0.58}$	$24.90_{\pm 0.96}$	126.8
GRPO $64 \rightarrow 64$	$21.45_{\pm 0.69}$	$35.20_{\pm 1.03}$	$15.75_{\pm 0.61}$	$27.80_{\pm 0.88}$	431.6
Random	$20.55_{\pm 0.82}$	$34.10_{\pm 1.18}$	$15.10_{\pm 0.67}$	$26.95_{\pm 1.01}$	298.7
Max-Reward	$20.10_{\pm 0.88}$	$33.45_{\pm 1.25}$	$14.85_{\pm 0.72}$	$26.40_{\pm 1.08}$	300.2
PODS	$21.85_{\pm 0.66}$	$35.95_{\pm 0.98}$	$16.10_{\pm 0.59}$	$28.55_{\pm 0.87}$	304.5
CPPO-Top-m	$21.50_{\pm 0.71}$	$35.60_{\pm 1.02}$	$15.90_{\pm 0.63}$	$28.20_{\pm 0.90}$	306.1
CBS*-quota	$22.10_{\pm 0.64}$	$36.30_{\pm 0.95}$	$16.30_{\pm 0.57}$	$28.80_{\pm 0.84}$	310.8
SSL (ours)	$\mathbf{22.95}_{\pm \mathbf{0.60}}$	$\mathbf{37.45}_{\pm \mathbf{0.91}}$	$\mathbf{17.05}_{\pm \mathbf{0.54}}$	$\mathbf{30.00}_{\pm \mathbf{0.80}}$	318.9
Qwen3-4B-Base					
GRPO $16 \rightarrow 16$	$22.80_{\pm 0.78}$	$37.20_{\pm 1.14}$	$17.10_{\pm 0.65}$	$29.90_{\pm 1.01}$	134.2
GRPO $64 \rightarrow 64$	$25.80_{\pm 0.71}$	$41.50_{\pm 1.02}$	$19.40_{\pm 0.60}$	$33.40_{\pm 0.90}$	458.7
Random	$24.90_{\pm 0.84}$	$40.30_{\pm 1.18}$	$18.80_{\pm 0.68}$	$32.40_{\pm 1.05}$	315.4
Max-Reward	$24.40_{\pm 0.91}$	$39.80_{\pm 1.24}$	$18.50_{\pm 0.73}$	$32.00_{\pm 1.10}$	316.8
PODS	$26.20_{\pm 0.68}$	$42.10_{\pm 0.96}$	$19.80_{\pm 0.58}$	$34.10_{\pm 0.86}$	321.2
CPPO-Top-m	$25.90_{\pm 0.73}$	$41.70_{\pm 1.00}$	$19.60_{\pm 0.61}$	$33.80_{\pm 0.89}$	323.0
CBS*-quota	$26.50_{\pm 0.65}$	$42.50_{\pm 0.92}$	$20.10_{\pm 0.56}$	$34.50_{\pm 0.82}$	327.5
SSL (ours)	$\mathbf{27.40}_{\pm \mathbf{0.61}}$	$\mathbf{43.80}_{\pm \mathbf{0.88}}$	$\mathbf{20.90}_{\pm \mathbf{0.52}}$	$\mathbf{35.70}_{\pm \mathbf{0.78}}$	336.1

Table 10: **Paired geometry, optimizer-transfer, and allocation comparisons.** Extended math data and the 3,072-token response cap are shared within each block. The GRPO/DAPO/GSPO panels use Qwen3-4B-Base; the Dr. GRPO panel uses Qwen2.5-Math-7B. Accuracy columns are mean ± SD over five seeds (%, ↑); Mean is the seven-benchmark average shown in Figure 2. Bold denotes the highest accuracy mean within each block.

	Avg@4	Avg@32				Training
Method	MATH500	AMC23	AIME24	AIME25	Mean	GPU-hours ↓
GRPO: geometry-based alternatives						
GRPO	$62.40_{\pm 0.79}$	$46.30_{\pm 1.20}$	$20.90_{\pm 1.15}$	$14.29_{\pm 1.04}$	$41.15_{\pm 0.58}$	320.0
+ PODS	$63.15_{\pm 0.76}$	$47.09_{\pm 1.16}$	$21.71_{\pm 1.08}$	$14.90_{\pm 0.98}$	$41.80_{\pm 0.60}$	320.0
+ SALT	$63.45_{\pm 0.73}$	$47.25_{\pm 1.12}$	$21.90_{\pm 1.05}$	$15.00_{\pm 0.96}$	$41.98_{\pm 0.56}$	320.0
+ GeoAlign	$63.70_{\pm 0.71}$	$47.45_{\pm 1.10}$	$22.10_{\pm 1.01}$	$15.25_{\pm 0.93}$	$42.16_{\pm 0.54}$	320.0
+ SSL (ours)	$\mathbf{64.20}_{\pm \mathbf{0.72}}$	$\mathbf{48.09}_{\pm \mathbf{1.12}}$	$\mathbf{22.79}_{\pm \mathbf{1.05}}$	$\mathbf{15.90}_{\pm \mathbf{0.94}}$	$\mathbf{42.56}_{\pm \mathbf{0.55}}$	320.0
DAPO: fixed-pool selector transfer						
+ Random	$64.00_{\pm 0.82}$	$47.80_{\pm 1.24}$	$22.60_{\pm 1.12}$	$15.50_{\pm 1.01}$	$42.37_{\pm 0.61}$	326.0
+ PODS	$65.00_{\pm 0.76}$	$48.70_{\pm 1.15}$	$23.50_{\pm 1.06}$	$16.40_{\pm 0.96}$	$43.14_{\pm 0.56}$	329.2
+ SSL (ours)	$\mathbf{66.00}_{\pm \mathbf{0.70}}$	$\mathbf{49.59}_{\pm \mathbf{1.10}}$	$\mathbf{24.40}_{\pm \mathbf{1.02}}$	$\mathbf{17.29}_{\pm \mathbf{0.91}}$	$\mathbf{43.89}_{\pm \mathbf{0.53}}$	338.5
GSPO: fixed-pool selector transfer						
+ Random	$64.80_{\pm 0.79}$	$48.41_{\pm 1.20}$	$23.21_{\pm 1.10}$	$16.10_{\pm 0.98}$	$42.91_{\pm 0.59}$	334.0
+ PODS	$65.70_{\pm 0.74}$	$49.20_{\pm 1.13}$	$24.00_{\pm 1.04}$	$16.90_{\pm 0.93}$	$43.61_{\pm 0.55}$	337.1
+ SSL (ours)	$\mathbf{66.70}_{\pm \mathbf{0.69}}$	$\mathbf{50.09}_{\pm \mathbf{1.07}}$	$\mathbf{24.90}_{\pm \mathbf{0.99}}$	$\mathbf{17.79}_{\pm \mathbf{0.89}}$	$\mathbf{44.36}_{\pm \mathbf{0.51}}$	346.0

Continued on next page

Table 10: Paired geometry, optimizer-transfer, and allocation comparisons (continued).

	Avg@4	Avg@32				Training
Method	MATH500	AMC23	AIME24	AIME25	Mean	GPU-hours ↓
Dr. GRPO: adaptive allocation						
Dr. GRPO	$70.80_{\pm 0.69}$	$57.91_{\pm 1.21}$	$30.21_{\pm 1.08}$	$21.29_{\pm 0.97}$	$48.84_{\pm 0.53}$	312.0
+ VIP	$71.60_{\pm 0.63}$	$58.80_{\pm 1.14}$	$31.10_{\pm 1.01}$	$22.21_{\pm 0.91}$	$49.59_{\pm 0.49}$	312.0
+ SSL (ours)	$\mathbf{72.20}_{\pm \mathbf{0.59}}$	$\mathbf{59.41}_{\pm \mathbf{1.08}}$	$\mathbf{31.90}_{\pm \mathbf{0.96}}$	$\mathbf{23.00}_{\pm \mathbf{0.87}}$	$\mathbf{50.16}_{\pm \mathbf{0.46}}$	312.0

References: SALT (Chang et al., 2026), GeoAlign (Zhou et al., 2026), DAPO (Yu et al., 2025), GSPO (Zheng et al., 2025), Dr. GRPO (Liu et al., 2025), and VIP (Nguyen et al., 2026). Geometry and allocation blocks compare native methods at 320 and 312 A100-GPU-hours, respectively, including auxiliary costs. DAPO/GSPO selector blocks use $64 \to 16$ at update 400 and retain each optimizer's loss and clipping conventions. VIP uses native adaptive allocation; its paired Dr. GRPO reference follows the same allocation-budget setting. SSL retains $64 \to 16$ and is evaluated at the same GPU-hour budget. Mean includes all seven benchmarks.

C.5 Paired normalization and representation controls

These controls use the shared frozen groups, eligibility rules, and timing protocol in Appendix B.6.

Normalization control. SSL has the highest mean cosine and lowest mean relative error under both normalizations. Compared with CBS*-quota, the strongest alternative on these metrics, its cosine increases by 0.05 in each condition and its relative error decreases by 0.08 and 0.07, respectively. Using full-group advantages raises SSL's cosine from 0.79 to 0.82, while its advantage over the comparators remains. The fidelity advantage therefore persists with reward normalization held fixed.

Table 11: **Selection versus advantage normalization.** Frozen Qwen3-4B-Base groups, $64 \to 16$. The same selected indices are reused in both conditions. Metrics compare independent parameter-gradient sketches; larger cosine and smaller relative error are better.

	Subset advantages $a_i(S)$		Full-group advantages $a_i(N)$	
Method	Cosine ↑	Rel. error ↓	Cosine ↑	Rel. error ↓
Random (Xu et al., 2025)	$0.63_{\pm 0.18}$	$0.88_{\pm 0.22}$	$0.70_{\pm 0.16}$	$0.78_{\pm 0.19}$
Max-Reward (Xu et al., 2025)	$0.52_{\pm 0.24}$	$0.98_{\pm 0.20}$	$0.64_{\pm 0.20}$	$0.87_{\pm 0.18}$
PODS (Xu et al., 2025)	$0.72_{\pm 0.16}$	$0.75_{\pm 0.19}$	$0.75_{\pm 0.15}$	$0.69_{\pm 0.17}$
CPPO-Top-m (Lin et al., 2025)	$0.69_{\pm 0.18}$	$0.81_{\pm 0.21}$	$0.73_{\pm 0.16}$	$0.74_{\pm 0.18}$
CBS*-quota (Lu et al., 2026)	$0.74_{\pm 0.15}$	$0.71_{\pm 0.18}$	$0.77_{\pm 0.14}$	$0.66_{\pm 0.16}$
SSL (ours)	$\mathbf{0.79}_{\pm \mathbf{0.13}}$	$\mathbf{0.63}_{\pm \mathbf{0.16}}$	$\mathbf{0.82}_{\pm \mathbf{0.12}}$	$\mathbf{0.59}_{\pm \mathbf{0.14}}$

Diagnostic coverage. Nonzero-gradient groups: 122/128; reference-zero groups: 6/128. The same eligibility mask applies to every selector and both normalization conditions.

Representation dimension. Increasing d from 32 to 128 improves cosine from 0.72 to 0.79 and reduces relative error from 0.75 to 0.63. The $d = 256$ representation reaches 0.80 cosine and 0.61 relative error, at 196.5 versus 118.4 ms per group (66% more selector time). The default balances this small fidelity difference against lower selection cost.

C.6 Permutation convergence and exact calibration

The large-group reference, exact small-group construction, and uncertainty metrics are defined in Appendix B.6.

Large-group convergence. At $K = 512$, rank correlation with the independent $K = 8192$ reference is 0.958 and selected-set overlap is 0.929. Increasing K to 2,048 raises overlap to 0.966

Table 12: **SSL representation fidelity and cost.** Only the pooling dimension changes. Gradient metrics use an independent, fixed 2,048-dimensional projection. Overlap is relative to $d = 128$ and ranges from zero to one.

		External gradient fidelity		Selection	Selector
Method	d	Cosine ↑	Rel. error ↓	Overlap	ms/group ↓
SSL	32	$0.72_{\pm 0.16}$	$0.75_{\pm 0.19}$	$0.73_{\pm 0.12}$	$58.9_{\pm 6.7}$
SSL	64	$0.76_{\pm 0.14}$	$0.68_{\pm 0.17}$	$0.84_{\pm 0.09}$	$79.7_{\pm 7.3}$
SSL (ours)	**128**	$\mathbf{0.79_{\pm 0.13}}$	$\mathbf{0.63_{\pm 0.16}}$	$\mathbf{1.000_{\pm 0.000}}$	$\mathbf{118.4_{\pm 8.2}}$
SSL	256	$0.80_{\pm 0.13}$	$0.61_{\pm 0.15}$	$0.91_{\pm 0.07}$	$196.5_{\pm 10.7}$

and external cosine from 0.790 to 0.796 while increasing latency from 118.4 to 446.8 ms per group (3.77×). The $K = 512$ cosine and latency means agree with the default rows in Tables 11 and 12. Their SDs differ because those tables summarize variation across groups, whereas Table 3 summarizes variation across seed-level group averages.

Exact small-group calibration. The attribution RMSE decreases from 0.0449 at $K = 32$ to 0.0056 at $K = 2048$, approximately an eightfold reduction for 64 times as many permutations. At $K = 512$, RMSE is 0.0112 and RMS-SE is 0.0114, with Spearman correlation 0.970 and overlap 0.939 against exact values. These measurements show how sampling accuracy improves with permutation count for the specified proxy game.

Table 13: **SSL attribution calibration against exact values.** $n = 16$, $m = 4$, $d = 128$. All sampled estimates and the exact reference use the same small-group game. RMSE is attribution error, distinct from the gradient relative error in Table 11.

		Attribution accuracy		Agreement with exact			Selector
Method	K	RMSE ↓	RMS-SE ↓	Spearman ↑	Overlap ↑	V_ρ	ms/group ↓
SSL (MC)	32	$0.0449_{\pm 0.0031}$	$0.0456_{\pm 0.0028}$	$0.782_{\pm 0.016}$	$0.744_{\pm 0.013}$	1117	$4.8_{\pm 0.2}$
SSL (MC)	64	$0.0318_{\pm 0.0022}$	$0.0323_{\pm 0.0020}$	$0.849_{\pm 0.012}$	$0.807_{\pm 0.010}$	1128	$6.7_{\pm 0.2}$
SSL (MC)	128	$0.0225_{\pm 0.0015}$	$0.0228_{\pm 0.0014}$	$0.904_{\pm 0.009}$	$0.863_{\pm 0.008}$	1135	$10.5_{\pm 0.3}$
SSL (MC)	256	$0.0158_{\pm 0.0011}$	$0.0161_{\pm 0.0010}$	$0.944_{\pm 0.006}$	$0.907_{\pm 0.006}$	1138	$18.3_{\pm 0.4}$
SSL (MC)	**512**	$\mathbf{0.0112_{\pm 0.0008}}$	$\mathbf{0.0114_{\pm 0.0007}}$	$\mathbf{0.970_{\pm 0.004}}$	$\mathbf{0.939_{\pm 0.005}}$	**1139**	$\mathbf{33.9_{\pm 0.6}}$
SSL (MC)	1,024	$0.0079_{\pm 0.0005}$	$0.0081_{\pm 0.0005}$	$0.985_{\pm 0.003}$	$0.961_{\pm 0.004}$	1140	$65.1_{\pm 1.0}$
SSL (MC)	2,048	$0.0056_{\pm 0.0004}$	$0.0057_{\pm 0.0003}$	$0.993_{\pm 0.001}$	$0.977_{\pm 0.003}$	1140	$127.8_{\pm 1.8}$
SSL (exact)	—	**0**	—	**1**	**1**	$114^\dagger$	—

Small-group reference coverage. Constant exact rankings: 14/128. †The exact row counts 114 distinct eligible groups, each evaluated once; MC rows count valid group–seed pairs, with a maximum of $114 \times 10 = 1140$. Exact-enumeration latency is unavailable.

C.7 Selected-set geometry and attribution

The cohort, selection rules, and density-estimation procedure are specified in Appendix B.7.

Table 14 gives the complete summaries. Relative to PODS, SSL increases mean novelty by 0.0892 (66.8%) and mean extremity by 5.4%, while reducing the selected positive-reward fraction by 8.01 percentage points. Percent changes use unrounded means. Its additional zero-reward rollouts have negative full-group advantages and contribute through their signed score directions.

Attribution consistency. Summing the raw estimates within each group recovers the efficiency target $v(N) - v(\varnothing)$, equal to one for eligible groups and zero for the excluded groups, with maximum absolute discrepancy 8.9×10^{-16} in the recorded selections. This checks the estimator's efficiency identity. The higher selected mean estimate for SSL (0.0575 versus 0.0267 for PODS)

Table 14: **Selected-set statistics for the geometry diagnostic.** Means over 1,984 selected rollouts per method from 124 eligible groups. Rewards are binary; $\widetilde{\phi}$ denotes the raw signed Shapley estimate.

Method	Extremity E	Novelty D	Reward r	$\widetilde{\phi}$
Random	0.8793	0.1206	0.5932	0.0161
Max-Reward	0.8414	0.0929	0.9889	0.0117
PODS	1.1281	0.1335	0.5181	0.0267
SSL	1.1895	0.2226	0.4380	0.0575

also follows from its ranking rule. Appendix C.6 evaluates attribution accuracy against exact values; the landscape describes the geometry of the selected rollouts.

Interpretation. Novelty occupies two separated bands, $D \leq 0.0518$ and $D \geq 0.4219$, so centroids between them need not represent observed rollouts. Positive lift in the upper band concentrates at intermediate extremity; its low-extremity edge includes negative lift, and tail bins have limited support. Novelty–attribution associations also vary at the lowest extremity. Since D_i compares each signed direction with its nearest individual neighbor, combinations of directions require a different measure. This descriptive analysis uses one selection per method; gradient fidelity is assessed separately in Appendix C.5.

C.8 Gradient fidelity by reward balance and conflict

Table 15 reports the unsmoothed results for the cohort and gradient response defined in Appendix B.8.

Table 15: **Gradient fidelity by reward balance and conflict.** Unsmoothed means over eligible prompt groups. The four corner regions use the stated strict thresholds and leave intermediate groups unassigned. The final row pools all eligible groups.

Region	Groups	PODS cosine	SSL cosine	Mean Δ_{fid}
$B_N < 0.40,\ C_N < 0.35$	219	0.7828	0.7800	-0.0028
$B_N > 0.75,\ C_N < 0.40$	145	0.7696	0.7986	0.0290
$B_N < 0.40,\ C_N > 0.65$	26	0.6926	0.7203	0.0277
$B_N > 0.75,\ C_N > 0.65$	744	0.6839	0.7930	0.1091
All eligible groups	2,483	0.7223	0.7844	0.0621

Aggregation. Table 15 reports equally weighted group means from unrounded paired differences. Per-seed gains are 0.0629, 0.0614, 0.0631, 0.0597, and 0.0632 (equal-seed mean 0.0621; sample SD 0.0015). The four regions contain 1,134 groups; the other 1,349 remain in the landscape and overall mean. Different low-conflict thresholds (0.35 and 0.40) make these unequal regions descriptive comparisons.

Interpretation. Mean gains are 0.0290 with high balance and low conflict, and 0.0277 with low balance and high conflict, versus 0.1091 when both are high. These are descriptive associations across groups nested within training runs; the low-balance, high-conflict region has only 26 groups. Smoothing estimates local mean gains; group density shows empirical support, with insufficient support marked gray.

D Proofs

This section proves Theorem 1, supplies the score-to-gradient bridge in Proposition 4, and establishes the associated sharpness and calibration results. All norms are Euclidean unless an operator norm is specified. The analysis is deterministic after conditioning on the sampled group and the selector's randomness.

D.1 FROZEN OBJECTIVE AND THE CLIPPING NEIGHBORHOOD

Fix a prompt, n sampled rollouts, and the sampling parameter $\theta_0 \in \mathbb{R}^p$. Let $N = \{1, \dots, n\}$ and let $S \subset N$ have size $m \geq 1$. For each rollout, $L_i \geq 1$ is the number of valid tokens; all sums below are over those tokens. The token history is the sampled history and remains fixed. For every nonempty $T \subseteq N$, define

$$\begin{aligned} \rho_{it}(\theta) &= \frac{\pi_\theta(y_{it} \mid x, y_{i,<t})}{\pi_{\theta_0}(y_{it} \mid x, y_{i,<t})}, \\ F_T(\theta) &= \frac{1}{|T|} \sum_{i \in T} \frac{a_i(T)}{L_i} \sum_{t=1}^{L_i} \rho_{it}(\theta) - \lambda_{\mathrm{KL}} \mathcal{K}_T(\theta), \\ h_i &= \frac{1}{L_i} \sum_{t=1}^{L_i} \nabla_\theta \log \pi_{\theta_0}(y_{it} \mid x, y_{i,<t}), \qquad k_T = \nabla \mathcal{K}_T(\theta_0). \end{aligned} \tag{16}$$

Here $a_i(T)$ is the ordinary, subset-recomputed advantage from equation 3, held constant when differentiating. The function $\mathcal{K}_T$ is the differentiable, token-averaged KL estimator used by the corresponding update, with all reference quantities fixed. Its dependence on T is retained when comparing gradients. Formula equation 16 describes the unclipped ascent surrogate. At the sampling parameters it agrees to first order with the negative clipped loss in equation 7.

For compactness, write $D_T = \sigma_T + \varepsilon_{\mathrm{adv}} > 0$ and introduce the zero-padded coefficient vector and matrices

$$\begin{aligned} w_{T,i} &= \frac{\mathbf{1}\{i \in T\}(r_i - \mu_T)}{|T| D_T}, \quad H = [h_1, \dots, h_n], \\ p_T &= \widetilde{g}(T) = Z^\top w_T, \qquad C = I_n - \frac{1}{n} \mathbf{1}\mathbf{1}^\top. \end{aligned} \tag{17}$$

The rows of Z are $z_i^\top$, so $p_T \in \mathbb{R}^d$ and $H \in \mathbb{R}^{p \times n}$. The matrix C is the orthogonal projection onto the centered subspace. In particular,

$$\mathbf{1}^\top w_T = \frac{\sum_{i \in T}(r_i - \mu_T)}{|T| D_T} = 0, \qquad C w_T = w_T, \qquad g_T := \nabla F_T(\theta_0) = H w_T - \lambda_{\mathrm{KL}} k_T. \tag{18}$$

Lemma 2 (Gradient identity and sufficient clipping condition)**.** *At θ_0, g_T in equation 18 is the negative gradient of the clipped loss. Let the clipping threshold be $\eta \in (0, 1)$. Suppose every valid token in N has a differentiable log probability whose gradient norm is at most M on the segment from θ_0 to $\theta_0 + \alpha g_S$, where $\alpha \geq 0$. If $\alpha M \|g_S\| < \log(1 + \eta)$, clipping is strictly inactive on that segment. With a non-strict inequality, the clipped and unclipped objective values still agree on the segment.*

Proof. All token likelihood ratios equal one at θ_0, strictly inside the clipping interval. For the finite collection of sampled tokens, continuity gives a neighborhood of θ_0 in which the two branches of each clipped minimum coincide with $\rho_{it}(\theta) a_i(T)$. The chain rule gives

$$\begin{aligned} \nabla \rho_{it}(\theta_0) &= \rho_{it}(\theta_0) \nabla \log \pi_{\theta_0}(y_{it} \mid x, y_{i,<t}) \\ &= \nabla \log \pi_{\theta_0}(y_{it} \mid x, y_{i,<t}), \\ -\nabla \mathcal{L}_{\mathrm{GRPO},T}(\theta_0) &= \frac{1}{|T|} \sum_{i \in T} a_i(T) h_i - \lambda_{\mathrm{KL}} k_T = g_T. \end{aligned} \tag{19}$$

For the finite-step claim, set $d = \alpha g_S$. For any $s \in [0, 1]$, the fundamental theorem of calculus and Cauchy–Schwarz imply

$$\begin{aligned} |\log \rho_{it}(\theta_0 + sd)| &= \left| \int_0^s \langle \nabla \log \pi_{\theta_0 + ud}(y_{it} \mid x, y_{i,<t}), d \rangle \, du \right| \\ &\leq \int_0^s |\langle \nabla \log \pi_{\theta_0 + ud}(y_{it} \mid x, y_{i,<t}), d \rangle| \, du \\ &\leq \int_0^s \|\nabla \log \pi_{\theta_0 + ud}(y_{it} \mid x, y_{i,<t})\| \, \|d\| \, du \\ &\leq s M \alpha \|g_S\| \leq M \alpha \|g_S\| \leq \log(1 + \eta). \end{aligned} \tag{20}$$

Exponentiation, followed by an elementary comparison, yields

$$1-\eta \le \frac{1}{1+\eta} \le \rho_{it}(\theta_0+sd) \le 1+\eta, \qquad \frac{1}{1+\eta}-(1-\eta)=\frac{\eta^2}{1+\eta}\ge 0. \tag{21}$$

Consequently clipping leaves every ratio unchanged, irrespective of the sign of the advantage. Under the strict assumption, all ratios lie strictly inside the interval. The KL term is identical in the clipped and unclipped objectives, which proves the assertion. □

For the remainder of this section, F_N has L-Lipschitz gradient on a convex neighborhood $\mathcal{U}$ of θ_0, where $L > 0$. A step is *admissible* if its segment lies in $\mathcal{U}$. The results always apply to the frozen unclipped objective. To interpret them as finite improvements in the clipped objective, we additionally require clipping to be inactive, for example by Lemma 2. Applying the calibrated step requires checking both smoothness and clipping admissibility in addition to proxy alignment.

D.2 Exact coefficient geometry under subset normalization

The overlap between w_S and w_N is available exactly, even though the two groups use different means and standard deviations. Retaining this overlap gives the joint discrepancy used below.

Lemma 3 (Centered coefficient identities)**.** *For any nonempty $S \subseteq N$ of size m and any scalar β, the vectors in equation 17 satisfy*

$$\begin{aligned} \|w_S\|^2 = \frac{\sigma_S^2}{mD_S^2}, \qquad \|w_N\|^2 = \frac{\sigma_N^2}{nD_N^2}, \qquad \langle w_S, w_N\rangle = \frac{\sigma_S^2}{nD_SD_N}, \\ \|w_S-\beta w_N\|^2 = \frac{\sigma_S^2}{mD_S^2}+\frac{\beta^2\sigma_N^2}{nD_N^2}-\frac{2\beta\sigma_S^2}{nD_SD_N}. \end{aligned} \tag{22}$$

These formulas also hold when either variance is zero.

Proof. By the definition of the population variance and the zero padding,

$$\begin{aligned} \|w_S\|^2 &= \sum_{i\in S}\frac{(r_i-\mu_S)^2}{m^2D_S^2} = \frac{m\sigma_S^2}{m^2D_S^2} = \frac{\sigma_S^2}{mD_S^2}, \\ \|w_N\|^2 &= \sum_{i\in N}\frac{(r_i-\mu_N)^2}{n^2D_N^2} = \frac{n\sigma_N^2}{n^2D_N^2} = \frac{\sigma_N^2}{nD_N^2}. \end{aligned} \tag{23}$$

For the mixed product, the mean shift disappears because the subset-centered rewards sum to zero:

$$\begin{aligned} \langle w_S, w_N\rangle &= \frac{1}{mnD_SD_N}\sum_{i\in S}(r_i-\mu_S)(r_i-\mu_N) \\ &= \frac{1}{mnD_SD_N}\sum_{i\in S}(r_i-\mu_S)\big[(r_i-\mu_S)+(\mu_S-\mu_N)\big] \\ &= \frac{\sum_{i\in S}(r_i-\mu_S)^2+(\mu_S-\mu_N)\sum_{i\in S}(r_i-\mu_S)}{mnD_SD_N} \\ &= \frac{m\sigma_S^2}{mnD_SD_N} = \frac{\sigma_S^2}{nD_SD_N}. \end{aligned} \tag{24}$$

Expanding the square $\|w_S-\beta w_N\|^2$ and substituting these three identities proves equation 22. Every denominator remains positive at zero variance, so the identities also hold in that case. □

For comparison with a bound that treats the two residuals separately, $\sigma_T/D_T \le 1$ gives the valid but potentially weaker chain

$$\begin{aligned} \|w_S-\beta w_N\| &\le \|w_S\|+\beta\|w_N\| \\ &= \frac{\sigma_S}{\sqrt{m}D_S}+\frac{\beta\sigma_N}{\sqrt{n}D_N} \le \frac{1}{\sqrt{m}}+\frac{\beta}{\sqrt{n}}, \qquad \beta>0. \end{aligned} \tag{25}$$

Our certificate uses the exact expression in equation 22, which retains the overlap discarded in equation 25. The factors $m^{-1/2}$ and $n^{-1/2}$ arise algebraically from normalization in this deterministic bound.

D.3 SHARED RESIDUALS AND JOINT UPDATE DISCREPANCY

The group-wide bridge assumption is that there exist a single $B \in \mathbb{R}^{p\times d}$, $b \in \mathbb{R}^p$, and $E \in \mathbb{R}^{p\times n}$ such that

$$H = BZ^\top + b\mathbf{1}^\top + E, \qquad \|EC\|_{\mathrm{op}} \le \tau, \qquad \tau \ge 0. \tag{26}$$

The map and error bound are common to S and N, allowing structured errors and dependence between selection and gradients. A useful certificate requires a sufficiently small, separately calibrated value of τ.

Proposition 4 (Joint update discrepancy)**.** *Under equation 26, write $u_T = Bp_T - \lambda_{\mathrm{KL}}k_T$. For every selected subset S and $\beta > 0$, define $q_\beta = w_S - \beta w_N$ and $\Delta_\beta = \|u_S - \beta u_N\| + \tau\|q_\beta\|$. Then $\|g_S - \beta g_N\| \le \Delta_\beta$. For fixed centers and coefficients, Δ_β is the exact largest discrepancy over $\|EC\|_{\mathrm{op}} \le \tau$. The residual contribution is governed by the difference of the two coefficient vectors.*

Proof of Proposition 4. By equation 18 and equation 26, for every nonempty T,

$$\begin{aligned} g_T &= (BZ^\top + b\mathbf{1}^\top + E)w_T - \lambda_{\mathrm{KL}}k_T \\ &= Bp_T + b(\mathbf{1}^\top w_T) + ECw_T - \lambda_{\mathrm{KL}}k_T \\ &= u_T + ECw_T, \qquad u_T := Bp_T - \lambda_{\mathrm{KL}}k_T. \end{aligned} \tag{27}$$

Set $q_\beta = w_S - \beta w_N$ and $v_\beta = u_S - \beta u_N$. Since $Cq_\beta = q_\beta$, subtraction before taking norms gives

$$\begin{aligned} \|g_S - \beta g_N\| &= \|v_\beta + ECq_\beta\| \\ &\le \|v_\beta\| + \|ECq_\beta\| \\ &\le \|v_\beta\| + \|EC\|_{\mathrm{op}}\|q_\beta\| \\ &\le \|u_S - \beta u_N\| + \tau\|w_S - \beta w_N\| = \Delta_\beta. \end{aligned} \tag{28}$$

The bound accounts for the unknown residual direction through its worst-case alignment with v_β.

To prove exactness, first suppose $q_\beta \neq 0$. Choose a unit vector $a = v_\beta/\|v_\beta\|$ if $v_\beta \neq 0$, and any unit vector in $\mathbb{R}^p$ otherwise. Define

$$E_* := \tau a \frac{q_\beta^\top}{\|q_\beta\|}. \qquad E_*C = E_*, \quad \|E_*C\|_{\mathrm{op}} = \tau, \quad E_*Cq_\beta = \tau\|q_\beta\|a. \tag{29}$$

The centering identity follows from $Cq_\beta = q_\beta$; the operator-norm identity follows from the norm of a rank-one matrix. If $v_\beta \neq 0$, its direction is a, so

$$\|v_\beta + E_*Cq_\beta\| = \big\|\big(\|v_\beta\| + \tau\|q_\beta\|\big)a\big\| = \|v_\beta\| + \tau\|q_\beta\|. \tag{30}$$

If $v_\beta = 0$, the same equality follows directly. If $q_\beta = 0$, every admissible residual gives discrepancy $\|v_\beta\| = \Delta_\beta$; $E_* = 0$ is sufficient. Thus the largest discrepancy in the stipulated residual class is exactly Δ_β in every case. □

Keeping the KL contribution inside the center norm can also matter. More explicitly,

$$\begin{aligned} \|u_S - \beta u_N\| &= \big\|B(p_S - \beta p_N) - \lambda_{\mathrm{KL}}(k_S - \beta k_N)\big\| \\ &\le \|B(p_S - \beta p_N)\| + \lambda_{\mathrm{KL}}\|k_S - \beta k_N\|, \end{aligned} \tag{31}$$

where $\lambda_{\mathrm{KL}} \ge 0$. The first line is the quantity used in our certificate. Even identical KL gradients cancel only when $\beta = 1$, or when that common gradient is zero.

Combining Proposition 4 with Theorem 1, $\Delta_\beta < \beta\|g_N\|$ is sufficient for positive progress at an admissible step of size $1/(L\beta)$. The bridge supplies the additional calibration needed to connect Shapley's score-space game to gradient progress. Appendix D.7 distinguishes the separate sharpness results from their composed sufficient bound.

D.4 SMOOTHNESS, THE EXACT ENVELOPE, AND PROOF OF THEOREM 1

Sharpness and step-size extension. Under the assumptions of Theorem 1, its guarantee is minimax sharp over L-smooth objectives and the stated discrepancy ball. More generally, the exact worst-case one-step certificate for an admissible scalar step size $\alpha > 0$ is $Q_\beta(\alpha, \delta)$ in equation 35. The calibrated size $1/(L\beta)$ uniquely maximizes this certificate over all positive scalar step sizes. The proof below establishes the main theorem together with these extensions; restricting the admissible step sizes may restrict which certificate can be attained by a feasible step.

We first establish the deterministic inequality without an error bound. For any admissible d, Lipschitz continuity of the gradient and the fundamental theorem of calculus yield

$$\begin{aligned} F_N(\theta_0 + d) - F_N(\theta_0) &= \int_0^1 \langle \nabla F_N(\theta_0 + sd), d\rangle \, ds \\ &= \langle g_N, d\rangle + \int_0^1 \langle \nabla F_N(\theta_0 + sd) - g_N, d\rangle \, ds \\ &\geq \langle g_N, d\rangle - \int_0^1 \|\nabla F_N(\theta_0 + sd) - g_N\| \|d\| \, ds \\ &\geq \langle g_N, d\rangle - \int_0^1 Ls\|d\|^2 \, ds \\ &= \langle g_N, d\rangle - \frac{L}{2}\|d\|^2. \end{aligned} \tag{32}$$

For $d = \alpha g_S$, completing the square provides the exact algebraic identity

$$\alpha\langle g_N, g_S\rangle - \frac{L\alpha^2}{2}\|g_S\|^2 = \frac{\|g_N\|^2 - \|L\alpha g_S - g_N\|^2}{2L}. \tag{33}$$

This identity combines angle and gradient magnitudes into a single vector discrepancy.

Proof of Theorem 1. Write $e = g_S - \beta g_N$, so $\|e\| \leq \delta$, and $G = \|g_N\|$. For any $\alpha > 0$ whose step is admissible, expanding the right side of equation 32 before applying Cauchy–Schwarz gives

$$\begin{aligned} F_N(\theta_0 + \alpha g_S) - F_N(\theta_0) &\geq \alpha\langle g_N, \beta g_N + e\rangle - \frac{L\alpha^2}{2}\|\beta g_N + e\|^2 \\ &= \alpha\beta G^2 + \alpha\langle g_N, e\rangle - \frac{L\alpha^2}{2}\left(\beta^2 G^2 + 2\beta\langle g_N, e\rangle + \|e\|^2\right) \\ &= \alpha\beta\left(1 - \frac{L\alpha\beta}{2}\right)G^2 + \alpha(1 - L\alpha\beta)\langle g_N, e\rangle - \frac{L\alpha^2}{2}\|e\|^2 \\ &\geq \alpha\beta\left(1 - \frac{L\alpha\beta}{2}\right)G^2 - \alpha|1 - L\alpha\beta|\,|\langle g_N, e\rangle| - \frac{L\alpha^2}{2}\|e\|^2 \\ &\geq \alpha\beta\left(1 - \frac{L\alpha\beta}{2}\right)G^2 - \alpha|1 - L\alpha\beta|G\|e\| - \frac{L\alpha^2}{2}\|e\|^2 \\ &\geq \alpha\beta\left(1 - \frac{L\alpha\beta}{2}\right)G^2 - \alpha|1 - L\alpha\beta|G\delta - \frac{L\alpha^2}{2}\delta^2. \end{aligned} \tag{34}$$

The last step uses the fact that both subtracted terms are nondecreasing in the nonnegative residual norm. Define the resulting envelope by

$$Q_\beta(\alpha, \delta) := \alpha\beta\left(1 - \frac{L\alpha\beta}{2}\right)G^2 - \alpha|1 - L\alpha\beta|G\delta - \frac{L\alpha^2}{2}\delta^2. \tag{35}$$

At $\alpha_* = (L\beta)^{-1}$ the cross term vanishes exactly. Hence

$$\begin{aligned} F_N(\theta_0 + \alpha_* g_S) - F_N(\theta_0) &\geq \frac{G^2}{2L} - \frac{\|e\|^2}{2L\beta^2} \\ &\geq \frac{G^2}{2L} - \frac{\delta^2}{2L\beta^2} = \frac{1 - \epsilon^2}{2L}G^2 > 0, \qquad \epsilon := \frac{\delta}{\beta G} < 1. \end{aligned} \tag{36}$$

We establish exactness of Q_β by constructing an attaining instance. Fix any nonzero $g \in \mathbb{R}^p$ with $\|g\| = G$ and consider the globally L-smooth function

$$F_*(\theta_0 + h) = F_*(\theta_0) + \langle g, h\rangle - \frac{L}{2}\|h\|^2, \qquad \nabla F_*(\theta_0) = g. \tag{37}$$

Its gradient has Lipschitz constant exactly L. For $1 - L\alpha\beta \neq 0$, select

$$e_* = -\operatorname{sign}(1 - L\alpha\beta)\frac{\delta}{G}g, \qquad g_{S,*} = \beta g + e_*. \tag{38}$$

Then $\|e_*\| = \delta$, and the signed mixed term in equation 34 is its exact lower extreme. Since equation 32 is an equality for F_*, direct substitution gives

$$F_*(\theta_0 + \alpha g_{S,*}) - F_*(\theta_0) = Q_\beta(\alpha, \delta). \tag{39}$$

If $1 - L\alpha\beta = 0$, take any e_* with norm δ; the mixed term vanishes and the same equality holds. For $\delta = 0$, choose $e_* = 0$. Thus every $\alpha > 0$ has an attaining instance in the abstract smooth-objective/discrepancy-ball class. The construction establishes sharpness within this abstract class; realizability by a rollout policy imposes additional structure.

Finally, let $x = L\alpha\beta > 0$. Normalizing equation 35 gives two quadratics:

$$\frac{L}{G^2}Q_\beta(\alpha, \delta) = x - \frac{x^2}{2} - \epsilon x|1 - x| - \frac{\epsilon^2 x^2}{2} = \begin{cases} x(1-\epsilon) - \frac{1}{2}x^2(1-\epsilon)^2, & 0 < x \le 1, \\ x(1+\epsilon) - \frac{1}{2}x^2(1+\epsilon)^2, & x \ge 1. \end{cases} \tag{40}$$

For $0 \le \epsilon < 1$, the first branch is strictly increasing on $(0, 1)$ and the second is strictly decreasing on $(1, \infty)$, because

$$\begin{aligned} (1-\epsilon)\big[1 - x(1-\epsilon)\big] > 0, \quad 0 < x < 1, \\ (1+\epsilon)\big[1 - x(1+\epsilon)\big] < 0, \quad x > 1. \end{aligned} \tag{41}$$

The branches agree at $x = 1$, with value $(1 - \epsilon^2)/2$. Therefore the unique maximizer is $x = 1$, equivalently $\alpha_* = (L\beta)^{-1}$. Together with the attaining construction, this proves the minimax and optimal-step assertions. □

If admissible steps are restricted to a closed positive interval, the optimal certificate is obtained by clipping $1/(L\beta)$ to that interval, by equation 41. The admissible interval must be verified for the objective and proposed step. The full-gradient benchmark follows by inserting $d = g_N/L$ into equation 32, when that step is admissible. Comparing it with equation 36 compares two guarantees; the true full-gradient improvement can be strictly larger than its smoothness lower bound.

D.5 Calibration, certificates, and the ideal cosine case

An inverse scaling $t = 1/\beta$ makes the useful calibration quantity convex. It also makes explicit which additional quantities, beyond the cached score proxy, a verifiable certificate requires.

Corollary 5 (Convex calibration with a shared residual)**.** *Under equation 26, define, for $t > 0$,*

$$\mathcal{D}(t) := \|tu_S - u_N\| + \tau\|tw_S - w_N\|, \qquad G_{\mathrm{lb}} := \|u_N\| - \tau\|w_N\|. \tag{42}$$

For every admissible step $\theta_t = \theta_0 + (t/L)g_S$,

$$F_N(\theta_t) - F_N(\theta_0) \ge \frac{G^2 - \mathcal{D}(t)^2}{2L}, \qquad G := \|g_N\|. \tag{43}$$

The function $\mathcal{D}$ is convex and continuous. On any nonempty compact interval $I \subset (0, \infty)$ of admissible t values, it attains a minimum; every such minimizer maximizes the right side of equation 43. If $G_{\mathrm{lb}} > 0$ and $\mathcal{D}(t) < G_{\mathrm{lb}}$, then the computable lower bound $(G_{\mathrm{lb}}^2 - \mathcal{D}(t)^2)/(2L)$ is strictly positive.

Proof. Using the common residual matrix exactly as in Proposition 4 gives

$$\begin{aligned} \|tg_S - g_N\| &= \|tu_S - u_N + EC(tw_S - w_N)\| \\ &\le \|tu_S - u_N\| + \|EC(tw_S - w_N)\| \\ &\le \|tu_S - u_N\| + \tau\|tw_S - w_N\| = \mathcal{D}(t). \end{aligned} \tag{44}$$

Substituting $\alpha = t/L$ into equation 33 proves equation 43, without needing a positive-progress condition. For $t_1, t_2 > 0$ and $s \in [0, 1]$, apply the triangle inequality twice to obtain

$$\begin{aligned}\mathcal{D}(st_1 + (1-s)t_2) &= \|s(t_1u_S - u_N) + (1-s)(t_2u_S - u_N)\| \\ &\quad + \tau\|s(t_1w_S - w_N) + (1-s)(t_2w_S - w_N)\| \\ &\le s\|t_1u_S - u_N\| + (1-s)\|t_2u_S - u_N\| \\ &\quad + s\tau\|t_1w_S - w_N\| + (1-s)\tau\|t_2w_S - w_N\| \\ &= s\mathcal{D}(t_1) + (1-s)\mathcal{D}(t_2).\end{aligned} \tag{45}$$

Continuity follows from continuity of norms. The extreme value theorem gives a minimizer on I; nonnegativity of $\mathcal{D}$ means that minimizing $\mathcal{D}$ also minimizes its square. Finally, the reverse triangle inequality implies

$$\begin{aligned}G = \|u_N + ECw_N\| &\ge \|u_N\| - \|ECw_N\| \\ &\ge \|u_N\| - \|EC\|_{\mathrm{op}}\|w_N\| \\ &\ge \|u_N\| - \tau\|w_N\| = G_{\mathrm{lb}} > 0, \\ \frac{G^2 - \mathcal{D}(t)^2}{2L} &\ge \frac{G_{\mathrm{lb}}^2 - \mathcal{D}(t)^2}{2L} > 0.\end{aligned} \tag{46}$$

The assumed positivity of G_{lb} justifies squaring the norm lower bound. □

Computing this diagnostic certificate requires a calibrated B, a valid residual bound τ, the KL gradients, and a valid smoothness constant. These inputs require separate calibration. SSL uses the original optimizer; minimizing $\mathcal{D}$ serves the theoretical analysis.

Corollary 6 (A sharp cosine interpretation in the ideal bridge)**.** *Suppose $\tau = 0$, $\lambda_{\mathrm{KL}} = 0$, and B is a scaled isometry on $\mathrm{span}\{p_S, p_N\}$: there is $s_B > 0$ such that $\|Bx\| = s_B\|x\|$ for every x in that span. Assume both proxy norms exceed $\varepsilon_{\cos}$ and $c = v(S) > 0$. If the step with $\alpha_* = c\|p_N\|/(L\|p_S\|)$ is admissible, then*

$$F_N(\theta_0 + \alpha_* g_S) - F_N(\theta_0) \ge \frac{c^2}{2L}\|g_N\|^2. \tag{47}$$

This is the best smoothness certificate over positive scalar multiples of g_S, and its constant is attained by a concave quadratic.

Proof. Polarization of the isometry assumption gives $\langle Bx, By\rangle = s_B^2\langle x, y\rangle$ on the span. Since $g_T = Bp_T$, let $A = \|g_S\| > 0$ and $G = \|g_N\| > 0$. Then

$$\langle g_S, g_N\rangle = cAG, \qquad A = s_B\|p_S\|, \qquad G = s_B\|p_N\|. \tag{48}$$

For every $t > 0$, expansion and completion of the square give

$$\begin{aligned}\|tg_S - g_N\|^2 &= t^2A^2 - 2tcAG + G^2 \\ &= A^2\left(t - \frac{cG}{A}\right)^2 + (1-c^2)G^2 \\ &\ge (1-c^2)G^2,\end{aligned} \tag{49}$$

with equality uniquely at $t_* = cG/A = c\|p_N\|/\|p_S\| > 0$. Using this exact residual in equation 33, with $\alpha_* = t_*/L$, yields

$$\begin{aligned}F_N(\theta_0 + \alpha_* g_S) - F_N(\theta_0) &\ge \frac{G^2 - \|t_*g_S - g_N\|^2}{2L} \\ &= \frac{G^2 - (1-c^2)G^2}{2L} = \frac{c^2G^2}{2L}.\end{aligned} \tag{50}$$

Equation 49 proves optimality among scalar steps. The quadratic equation 37 attains the smoothness inequality for these fixed gradient vectors, proving sharpness in this ideal geometric class. Equivalently, Theorem 1 applies with $\beta_* = A/(cG)$ and relative error $\epsilon = \sqrt{1-c^2}$. The appropriate calibration minimizes the *normalized* discrepancy $\|g_S/\beta - g_N\|$, accounting for the scale dependence of the unnormalized discrepancy $\|g_S - \beta g_N\|$. □

D.6 SCORE AND GRADIENT FIDELITY IN AN EXPLICIT POLICY FAMILY

The following policy construction derives both log probabilities and their gradients, then applies exact Shapley selection with the prescribed tie-breaking rule.

Proposition 7 (Perfect proxy cosine can select an ascent failure)**.** *There exists a one-token, three-rollout group with binary rewards and $m = 2$ for which exact Shapley selection returns $S = \{1, 2\}$ with $v(S) = 1$, but $\langle g_N, g_S \rangle < 0$. With no KL penalty, every sufficiently small positive gradient step along g_S decreases the full-group surrogate. There is also a second policy family with the same frozen rollout probabilities, proxy game, rewards, Shapley values, and full-group gradient, for which the selected gradient has the opposite sign.*

Proof. Consider a scalar parameter θ and a policy over three possible one-token outputs. Let $\theta_0 = 0$ and define

$$\pi_\theta(i) = \frac{p_i \exp(\theta b_i)}{\sum_{j=1}^{3} p_j \exp(\theta b_j)}, \qquad (p_1, p_2, p_3) = \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2}\right), \qquad (r_1, r_2, r_3) = (0, 1, 1). \tag{51}$$

Condition on drawing these three distinct outputs in this order, an event with probability $p_1 p_2 p_3 > 0$ under independent sampling. Take the scalar projection $P = 1$, so $z_i = \log p_i$, and let $\varepsilon = \varepsilon_{\mathrm{adv}} > 0$. The two unequal-reward pairs have standard deviation $1/2$, and the full group has standard deviation $\sqrt{2}/3$. Setting $D_2 = 1/2 + \varepsilon$ and $D_3 = \sqrt{2}/3 + \varepsilon$, direct substitution gives

$$\begin{aligned} p_{\{1,2\}} &= \frac{\log p_2 - \log p_1}{4D_2} = \frac{\log 2}{4D_2} > 0, \\ p_{\{1,3\}} &= \frac{\log p_3 - \log p_1}{4D_2} = \frac{\log 3}{4D_2} > 0, \\ p_N &= \frac{-2\log p_1 + \log p_2 + \log p_3}{9D_3} = \frac{\log 6}{9D_3} > 0. \end{aligned} \tag{52}$$

Choose $\varepsilon_{\cos}$ strictly smaller than all three displayed positive values. Every singleton and the pair $\{2, 3\}$ have constant rewards and therefore zero proxy. Since the nonzero proxies are positive scalars, the entire coalition game is

$$v(\varnothing) = v(\{1\}) = v(\{2\}) = v(\{3\}) = v(\{2,3\}) = 0, \qquad v(\{1,2\}) = v(\{1,3\}) = v(N) = 1. \tag{53}$$

For $n = 3$, the Shapley weight is $1/3$ for coalitions of size zero or two and $1/6$ for coalitions of size one. Enumerating the nonzero marginal contributions yields

$$\begin{aligned} \phi_1 &= \tfrac{1}{6}[v(\{1,2\}) - v(\{2\})] + \tfrac{1}{6}[v(\{1,3\}) - v(\{3\})] + \tfrac{1}{3}[v(N) - v(\{2,3\})] = \tfrac{2}{3}, \\ \phi_2 &= \tfrac{1}{6}[v(\{1,2\}) - v(\{1\})] + \tfrac{1}{6}[v(\{2,3\}) - v(\{3\})] + \tfrac{1}{3}[v(N) - v(\{1,3\})] = \tfrac{1}{6}, \\ \phi_3 &= \tfrac{1}{6}. \end{aligned} \tag{54}$$

The omitted empty-coalition increments are zero; the last equality follows from symmetry of players two and three in equation 53. Thus the increasing-index tie-break selects $S = \{1, 2\}$, which has utility one.

Now take $(b_1, b_2, b_3) = (0, -1, 3)$ and $\lambda_{\mathrm{KL}} = 0$. Differentiating the actual policy in equation 51 gives

$$h_i = \frac{d}{d\theta} \log \pi_\theta(i)\Big|_{\theta=0} = b_i - \bar{b}, \qquad \bar{b} = \sum_{j=1}^{3} p_j b_j. \tag{55}$$

The common mean $\bar{b}$ cancels from the centered update. Consequently,

$$\begin{aligned} g_S &= \frac{h_2 - h_1}{4D_2} = \frac{b_2 - b_1}{4D_2} = -\frac{1}{4D_2} < 0, \\ g_N &= \frac{-2h_1 + h_2 + h_3}{9D_3} = \frac{-2b_1 + b_2 + b_3}{9D_3} = \frac{2}{9D_3} > 0, \\ g_N g_S &= -\frac{1}{18 D_2 D_3} < 0. \end{aligned} \tag{56}$$

All likelihoods are strictly positive and analytic near zero. The frozen full-group objective is therefore differentiable and locally smooth. By its directional derivative,

$$\lim_{\alpha \downarrow 0} \frac{F_N(\alpha g_S) - F_N(0)}{\alpha} = g_N g_S < 0. \tag{57}$$

For a direct strict inequality, differentiability supplies an $\alpha_0 > 0$ for which the Taylor remainder has absolute value at most $-\alpha g_N g_S/2$ whenever $0 < \alpha < \alpha_0$. Thus

$$F_N(\alpha g_S) - F_N(0) \le \alpha g_N g_S + \frac{-\alpha g_N g_S}{2} = \frac{\alpha g_N g_S}{2} < 0, \qquad 0 < \alpha < \alpha_0. \tag{58}$$

Continuity of the finitely many likelihood ratios, each equal to one at zero, permits decreasing α_0 so that clipping is inactive throughout these steps.

Finally, replace b by $b^+ = (0, 1, 1)$ in equation 51. At $\theta_0 = 0$ this leaves the probabilities, rewards, score vectors, utilities, and Shapley values unchanged. Repeating the centered derivative calculation gives

$$g_S^+ = \frac{1}{4D_2} = -g_S > 0, \qquad g_N^+ = \frac{2}{9D_3} = g_N > 0. \tag{59}$$

The two policy families share the same proxy observations and full-group gradient while producing selected gradients of opposite signs. Thus a connection from proxy fidelity to first-order ascent requires information about parameter-gradient geometry. □

D.7 Conditions for applying the guarantees

Sharpness of the component bounds. Proposition 4 is exact when the centers, coefficient vectors, and β are fixed and EC ranges over its operator-norm ball. Theorem 1 is exact when g_N is fixed and $g_S - \beta g_N$ ranges over its Euclidean ball, with the objective ranging over smooth functions. The two results optimize over different uncertainty sets. In the shared bridge model, changing E couples g_N with the discrepancy. This coupling can prevent simultaneous attainment of the separate extrema for fixed rollout data. Substituting Δ_β therefore gives a sufficient rollout guarantee with possible slack; replacing the full-gradient norm by G_{lb} can add further slack.

Selection and sampling. The argument applies pathwise to the realized subset, whether obtained by exact Shapley values, permutation sampling, or another selector, and allows dependence between selection and gradients. Optimality of top-m attribution for $v(S)$ or $\mathcal{D}(t)$, and comparisons with uniform selection, are separate questions. Probabilistic guarantees require a high-probability bridge event and, where relevant, an analysis of ranking error.

Normalization and optimizer. Centering is essential to the stated bridge reduction. For the fixed-normalization control, the selected coefficients are $w_{S,i}^{\mathrm{fix}} = \mathbf{1}\{i \in S\} a_i(N)/m$, whose sum is generally nonzero. The identity replacing equation 27 is

$$g_S^{\mathrm{fix}} = BZ^\top w_S^{\mathrm{fix}} + b(\mathbf{1}^\top w_S^{\mathrm{fix}}) + E w_S^{\mathrm{fix}} - \lambda_{\mathrm{KL}} k_S. \tag{60}$$

The uncentered residual term requires an additional bound or a different centering decomposition beyond the condition on EC. The smoothness identity remains valid for any actual direction. For Adam or a momentum optimizer, applying it requires the implemented direction, which depends on optimizer state and generally differs from g_S.

Objective and horizon. The result quantifies local progress from a small calibrated joint discrepancy for a single step on a frozen empirical full-group surrogate. Extending it to multiple steps requires bounds along the trajectory for policy ratios, gradients, normalization, and selection. Relating this progress to population return or verifier accuracy requires a further statistical or task-level analysis.