Design Conditions for Intra-Group Learning of Sequence-Level Rewards: Token Gradient Cancellation

Fei Ding [1] **Yongkang Zhang** [1] **Youwei Wang** [2] **Zijian Zeng** [2]
[1]Alibaba Group [2]Tsinghua University

Abstract

In sparse termination rewards, intra-group comparisons have become the dominant paradigm for fine-tuning reasoning models via reinforcement learning. However, long-term training often leads to issues like ineffective update accumulation (learning tax), solution probability drift, and entropy collapse. This paper presents a necessary condition for algorithm design from a token-level credit assignment perspective: to prevent reward-irrelevant drift, intra-group objectives must maintain gradient exchangeability across token updates, enabling gradient cancellation on weak-credit/high-frequency tokens. We show that two common mechanisms disrupting exchangeability make "non-cancellation" a structural norm. Based on this, we propose minimal intra-group transformations to restore or approximate the cancellation structure in the shared token space. Experimental results demonstrate that these transformations stabilize training, improve sample efficiency, and enhance final performance, validating the value of this design condition.

1. INTRODUCTION

Under sparse termination feedback, large language models (LLMs) have shown significant improvements in performance on complex reasoning tasks through reinforcement learning. Learning objectives based on intra-group comparisons have gradually become the mainstream paradigm. The core idea is to compare multiple candidate trajectories for the same input and learn through the intra-group relative relationships. However, while these objectives significantly improve performance early in training, long-term training struggles to maintain stability, resulting in phenomena such as ineffective update accumulation, equivalent solution probability drift, and entropy collapse.

Existing works typically attribute instability to reward sparsity or optimization noise. However, these explanations do not address a fundamental question: why do different intra-group learning objectives, despite differing implementation details, repeatedly exhibit similar failure modes? We argue that this phenomenon may arise from a structural limitation, rather than from specific algorithms or hyperparameter choices.

This paper presents a unified perspective from the viewpoint of token-level credit assignment: if intra-group objectives disrupt the exchangeability of token updates at the gradient level—especially during sequence-coupled trajectory aggregation or asymmetric segment pruning/selection—systematic drift (learning tax) and probability drift/entropy collapse are inevitable. We further demonstrate that when this condition is violated, the accumulation of learning tax and probability drift is generally predictable.

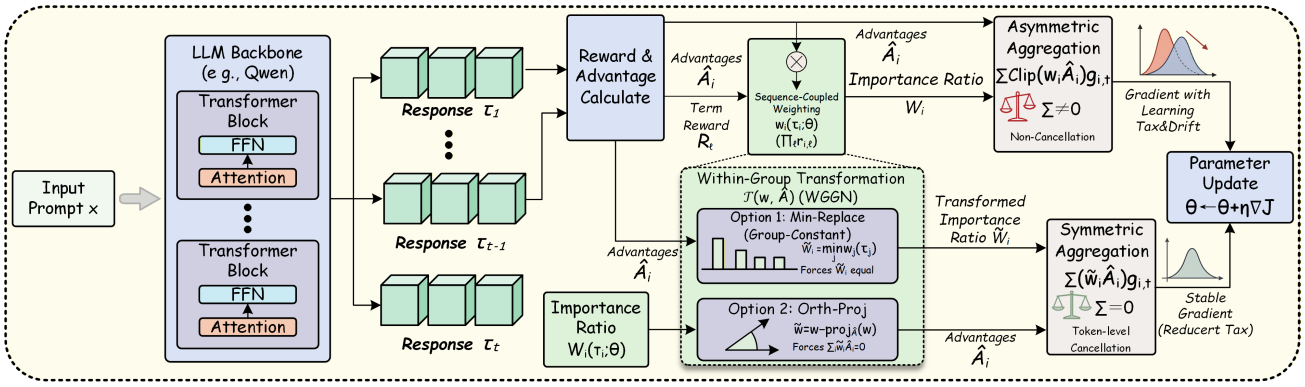
Contributions. The contributions of this paper include:

- **Structural Boundaries of Intra-Group Learning:** We propose a necessary condition that maintaining token-level gradient exchangeability is essential for stable intra-group learning under sparse termination feedback.
- **A Unified Gradient Perspective:** Through gradient analysis, we clearly distinguish between the behaviors of exchangeable and non-exchangeable objectives.
- **Constructive Validation of Structural Fixes:** We propose several minimal intra-group transformations that restore or approximate the gradient cancellation structure without altering the core framework of intra-group comparisons.

2. Related Work

Intra-Group Comparison-based Reinforcement Learning Objectives. Representative methods include GRPO (Shao et al., 2024), GSPO (Zheng et al., 2025), and their variants (e.g., DAPO (Yu et al., 2025), DCPO (Yang et al., 2025c), SSPO (Yang et al., 2025b)), which construct learning signals by comparing multiple trajectories for the same input and have shown advantages in tasks like

Correspondence to: Fei Ding <dignfei@gmail.com>.



mathematical reasoning. Unlike existing work, this paper reveals the structural boundaries of intra-group learning objectives from the perspective of token-level credit assignment, providing a unified explanation for the failure modes across different intra-group learning methods.

3. Necessary Conditions for Stable Intra-Group Comparison Learning Design

In intra-group comparison learning, when the objective function disrupts the token-level update commutativity in the gradient domain, the inherent compensatory structure within the group fails, leading to accumulated learning taxes and probability drift/entropy collapse. Specifically, we propose the following verifiable premise: there exists a class of token types weakly related to the reward (e.g., generic template tokens), and the learning objective follows the intra-group comparison paradigm. If the effective weights of these weakly related tokens lack commutativity within the group, continuous non-zero drift arises, ultimately resulting in systematic learning taxes and entropy collapse.

3.1. General Form: Unified Representation of Intra-Group Comparison Objectives

Consider multiple trajectories sampled from the reference policy for the same input, with a unified objective function given by:

$$\mathcal{J}(\theta) = \mathbb{E}_x \left[\frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{T_i} \omega_{i,t}(\tau_i;\theta) \widehat{A}_i \right], \tag{1}$$

where $\omega_{i,t}(\tau_i;\theta)$ represents the effective weight of the t-th token in the i-th trajectory, and $\widehat{A}_i$ denotes the intra-group comparison signal.

3.2. Definition of Intra-Group Cancellation

Intra-group cancellation refers to a zero-gradient contribution at the shared token for that timestep:

$$\frac{1}{G} \sum_{i=1}^{G} \widehat{A}_i \nabla_\theta \omega_{i,t^\star} = \mathbf{0}. \tag{2}$$

This means that if a token carries no credit information distinguishing the quality of trajectories, updating it results in ineffective updates, leading to entropy collapse.

3.3. Limitations of Statistical Cancellation

In practice, strict token-wise cancellation is rare when contexts are not perfectly consistent, but when weights are exchangeable, updates to generic tokens tend to cancel out across groups. However, when weights are non-exchangeable, the effective coefficients of generic tokens are biased by trajectory-level random factors, potentially resulting in significant net updates within the group. Hence, intra-group cancellation remains necessary.

Proposition 3.1 (Ineffective Updates and Distribution Drift without Intra-Group Cancellation)**.** *Fix an input x and timestep $t^\star$. Consider event $\mathcal{E}_{t^\star}$: all trajectories in the group share the same "context-token" pair $(h^\star, y^\star)$ at this timestep. Define the group aggregate gradient induced by timestep $t^\star$ as*

$$g_{t^\star} \triangleq \frac{1}{G} \sum_{i=1}^{G} \widehat{A}_i \, \nabla_\theta \omega_{i,t^\star}(\tau_i;\theta),$$

and let an update be $\theta^+ = \theta + \eta g_{t^\star}$ ($\eta > 0$ is the step size). Define the conditional Fisher information matrix at context $h^\star$ as

$$F_\theta(h^\star) \triangleq \mathbb{E}_{y \sim \pi_\theta(\cdot|h^\star)} \Big[\nabla_\theta \log \pi_\theta(y \mid h^\star) \, \nabla_\theta \log \pi_\theta(y \mid h^\star)^\top \Big].$$

If the event $\mathcal{E}_{t^\star}$ does not *satisfy intra-group cancellation, i.e., $g_{t^\star} \neq \mathbf{0}$, and $F_\theta(h^\star)$ is non-degenerate in the direction*

of $g_{t^\star}$:

$$g_{t^\star}^\top F_\theta(h^\star)\, g_{t^\star} \;>\; 0,$$

then this step will induce a strictly positive conditional distribution drift at context $h^\star$:

$$\begin{aligned}\mathrm{KL}\big(\pi_{\theta^+}(\cdot \mid h^\star) \,\big\|\, \pi_\theta(\cdot \mid h^\star)\big) &= \frac{1}{2}\eta^2\, g_{t^\star}^\top F_\theta(h^\star)\, g_{t^\star} \\ &+ O(\eta^3 \|g_{t^\star}\|^3) \quad > 0, \qquad (\eta \textit{ sufficiently small}).\end{aligned} \tag{3}$$

Thus, when a shared token carries no distinguishing credit information for trajectory quality, violating intra-group cancellation leads to ineffective updates *(reward-irrelevant drift) for that token distribution at context $h^\star$.*

Proof. Let $\delta\theta \triangleq \theta^+ - \theta = \eta g_{t^\star}$. For fixed context $h^\star$, consider the function

$$\varphi(\delta\theta) \triangleq \mathrm{KL}\big(\pi_{\theta+\delta\theta}(\cdot \mid h^\star) \,\big\|\, \pi_\theta(\cdot \mid h^\star)\big)\,.$$

We have $\varphi(\mathbf{0}) = 0$, and since KL achieves a minimum at identical distributions, $\nabla_{\delta\theta}\varphi(\mathbf{0}) = \mathbf{0}$. Performing a second-order Taylor expansion around $\delta\theta = \mathbf{0}$ gives

$$\varphi(\delta\theta) = \frac{1}{2}\,\delta\theta^\top\, \nabla^2_{\delta\theta}\varphi(\delta\theta)\big|_{\delta\theta=\mathbf{0}}\,\delta\theta + O(\|\delta\theta\|^3). \tag{4}$$

The standard result shows that the Hessian at $\delta\theta = \mathbf{0}$ equals the conditional Fisher information matrix:

$$\nabla^2_{\delta\theta}\varphi(\delta\theta)\big|_{\delta\theta=\mathbf{0}} = F_\theta(h^\star).$$

Substituting $\delta\theta = \eta g_{t^\star}$ yields the second-order expansion in (3). When $g_{t^\star} \neq \mathbf{0}$ and $g_{t^\star}^\top F_\theta(h^\star) g_{t^\star} > 0$, the leading term is strictly positive; choose a sufficiently small η to ensure that the third-order remainder does not change the sign, resulting in a strictly positive KL divergence and ineffective updates. □

Corollary 3.2 (Log-Odds Drift in the Equivalent Solution Set Accumulates ⇒ Entropy Collapse Trend (Linear Region))**.** *Fix input x, and consider two semantically equivalent outputs with the same reward, y_a, y_b. For a class of* sequence-coupled *intra-group objective linear segment updates (with stop-gradient on sequence effective coefficients):*

$$\theta^+ = \theta + \eta\,\kappa(y;\theta)\,\mathbf{g}(y;\theta), \qquad \kappa(y;\theta) \geq 0,$$

there exists a constant $c(x) > 0$ such that

$$\Delta \log \frac{\pi_\theta(y_a \mid x)}{\pi_\theta(y_b \mid x)} = \eta\, c(x)\big(\kappa(y_a;\theta) - \kappa(y_b;\theta)\big) + O(\eta^2). \tag{5}$$

If $\kappa(y_a;\theta_k) - \kappa(y_b;\theta_k)$ maintains the same sign with a lower bound $|\kappa(y_a;\theta_k) - \kappa(y_b;\theta_k)| \geq \underline{\Delta} > 0$ across K consecutive updates, the log-odds drift accumulates over K steps, causing $\pi_{\theta_K}(y_a \mid x)/\pi_{\theta_K}(y_b \mid x) \to 0$ or ∞, leading to entropy tending to 0 on the binary equivalent subset $\{y_a, y_b\}$, manifesting as an entropy collapse trend in the equivalent solution set.

Proof. For fixed y, the first-order Taylor expansion gives $\Delta \log \pi_\theta(y \mid x) = \eta\, c(x)\kappa(y;\theta) + O(\eta^2)$ (with $c(x) > 0$ absorbing normalization constants), and subtracting the two gives (5). Summing over $k = 0, \ldots, K-1$ gives the linear accumulation of log-odds; when its absolute value tends to infinity, the binary normalized probability $p_k = \frac{\pi_{\theta_k}(y_a|x)}{\pi_{\theta_k}(y_a|x)+\pi_{\theta_k}(y_b|x)}$ satisfies $p_k \to 0$ or 1, so the binary entropy $h(p_k) \to 0$, yielding the conclusion. □

A minimal algebraic example corresponding to the above conjecture is provided in Appendix A.

Corollary 3.3 (Sequence Coupling ⇒ Group-wise Cancellation Holds Only on Zero Measure Sets)**.** *For the event $\mathcal{E}_{t^\star}$ (groups share the same context-token pair $(h^\star, y^\star)$), if the effective weights exhibit a* multiplicative sequence coupling *structure in the time dimension, i.e.,*

$$\omega_{i,t^\star}(\tau_i;\theta) = r_{i,t^\star}(\theta)\cdot u_i(\theta), \qquad u_i(\theta) \triangleq \prod_{t\neq t^\star} r_{i,t}(\theta),$$

then the group-wise cancellation condition for shared tokens is

$$\sum_{i=1}^{G} \widehat{A}_i \nabla_\theta \omega_{i,t^\star}(\tau_i;\theta) = \mathbf{0},$$

which generally requires $u_1(\theta) = \cdots = u_G(\theta)$ (except in degenerate cases). This constraint set defined by the continuous parameter θ is generally of measure zero*, meaning "non-cancellation" is the structural norm under sequence-coupling weighting.*

Corollary 3.4 (Group Orthogonality/Alignment Does Not Force Overall Gradient to Zero)**.** *Assume we construct weights $\widetilde{\mathbf{s}}$ within each group such that*

$$\sum_{i=1}^{G} \widetilde{s}_i \widehat{A}_i = 0,$$

and apply stop-gradient to $\widetilde{\mathbf{s}}$ during backpropagation (treating it as a constant coefficient). The above constraint only *restores (or approximates) cancellation structure in the "shared/high-frequency token subspace where gradients are highly aligned within the group," but* does not generally imply *the overall policy gradient is zero, because $\nabla_\theta \log \pi_\theta(\cdot)$ directions are not aligned across different trajectories/time-steps. Full decomposition and formal discussion can be found in Appendix D.*

3.4. Example: Comparison of Cancellation and Non-Cancellation

Using GRPO and GSPO as examples, we omit other algorithms due to space limitations. To highlight structural differences, we analyze within the linear region of the objective function (ignoring piecewise effects of $\min/\mathrm{clip}$; this

local analysis does not alter the conclusions about gradient decomposability/coupling). Consider a group of trajectories τ_1, τ_2 with size $G = 2$ under the same input x, where the group advantage satisfies

$$\widehat{A}_1 = -\widehat{A}_2 \triangleq -A, \qquad A > 0. \tag{6}$$

Let the token-level importance ratio be

$$r_{i,t}(\theta) = \frac{\pi_\theta(a_t^{(i)} \mid h_t^{(i)})}{\pi_{\theta_{\text{old}}}(a_t^{(i)} \mid h_t^{(i)})}. \tag{7}$$

GRPO (Token-Factorized). In the token-factorized form of GRPO (linear segment), the objective can be written as

$$\mathcal{J}_{\text{tok}}(\theta) = \frac{1}{2} \sum_{i=1}^{2} \sum_{t} r_{i,t}(\theta) \widehat{A}_i. \tag{8}$$

For a shared context-token pair $(h^\star, a^\star)$, assume that at time-step $t^\star$, both trajectories satisfy $h_{t^\star}^{(1)} = h_{t^\star}^{(2)} = h^\star$ and $a_{t^\star}^{(1)} = a_{t^\star}^{(2)} = a^\star$, and within the local neighborhood, $r_{1,t^\star}(\theta) = r_{2,t^\star}(\theta) = \rho$. The gradient contribution for this token pair is

$$\begin{aligned} \nabla_\theta \mathcal{J}_{\text{tok}}^{(t^\star)} &= \frac{1}{2} \left(\widehat{A}_1 \nabla_\theta r_{1,t^\star}(\theta) + \widehat{A}_2 \nabla_\theta r_{2,t^\star}(\theta) \right) \\ &= \frac{1}{2} \left(\widehat{A}_1 \rho + \widehat{A}_2 \rho \right) \nabla_\theta \log \pi_\theta(a^\star \mid h^\star) \\ &= \frac{\rho}{2} (\widehat{A}_1 + \widehat{A}_2) \nabla_\theta \log \pi_\theta(a^\star \mid h^\star) = \mathbf{0}, \end{aligned} \tag{9}$$

where the last step uses $\widehat{A}_1 + \widehat{A}_2 = 0$. Thus, *the gradient of the shared context-token cancels within the group.*

Asymmetric Clipping Breaks Cancellation (Minimal Illustration). The strict cancellation above relies on the key assumption that shared tokens are multiplied by *the same effective coefficient* within the group. However, in practice, many methods introduce piecewise operators (e.g., $\min$ /clip or threshold-based selection), which modify the effective weights for each trajectory as

$$\tilde{\omega}_{i,t}(\theta) = \phi_i \left(r_{i,t}(\theta) \right), \tag{10}$$

where $\phi_i(\cdot)$ is a *group-inconsistent* piecewise mapping induced by advantage signs, threshold triggers, or implementation details. Even if $r_{1,t^\star} = r_{2,t^\star}$ for a shared context-token event, as long as $\phi_1 \neq \phi_2$ (or the piecewise intervals trigger differently), we have $\nabla_\theta \tilde{\omega}_{1,t^\star} \neq \nabla_\theta \tilde{\omega}_{2,t^\star}$, so that $\widehat{A}_1 + \widehat{A}_2 = 0$ no longer guarantees cancellation. Therefore, *asymmetric piecewise clipping can systematically break exchangeability-cancellation even under token-factorized structures*.

In objectives with piecewise operators like $\min$ /clip, the exchangeability assumption is often violated, so cancellation no longer holds; a complete derivation for typical asymmetric clipping scenarios is provided in Appendix F.

GSPO (sequence-coupled). In the sequence-coupled form of GSPO (linear segment), the objective is

$$\mathcal{J}_{\text{seq}}(\theta) = \frac{1}{2} \sum_{i=1}^{2} s_i(\theta) \, \widehat{A}_i, \qquad s_i(\theta) = \prod_t r_{i,t}(\theta). \tag{11}$$

For shared time step $t^\star$, we decompose s_i as

$$s_i(\theta) = r_{i,t^\star}(\theta) \cdot u_i(\theta), \qquad u_i(\theta) \triangleq \prod_{t \neq t^\star} r_{i,t}(\theta). \tag{12}$$

The gradient term aligned with $\nabla_\theta \log \pi_\theta(a^\star \mid h^\star)$ is

$$\begin{aligned} \nabla_\theta \mathcal{J}_{\text{seq}}^{(t^\star)} &= \frac{1}{2} \left(\widehat{A}_1 s_1(\theta) + \widehat{A}_2 s_2(\theta) \right) \nabla_\theta \log \pi_\theta(a^\star \mid h^\star) + \cdots \\ &= \frac{1}{2} \left(-A \cdot \rho u_1(\theta) + A \cdot \rho u_2(\theta) \right) \nabla_\theta \log \pi_\theta(a^\star \mid h^\star) + \cdots \\ &= \frac{A\rho}{2} \left(u_2(\theta) - u_1(\theta) \right) \nabla_\theta \log \pi_\theta(a^\star \mid h^\star) + \cdots . \end{aligned} \tag{13}$$

This term is strictly nonzero as long as $u_1(\theta) \neq u_2(\theta)$, meaning *the gradients of shared context-tokens cannot cancel within the group*.

Structural Explanation. Equation (13) indicates that cancellation only occurs in the degenerate set where $u_1(\theta) = u_2(\theta)$; this set is defined by strict equality constraints and has measure zero in the continuous parameter space. Hence, intra-group non-cancellation of shared tokens is the structural norm for sequence-coupled weighting.

4. Method: Decoupled Group-relative Gradient Estimator Aligned with Structural Proposition

Instead of starting from a specific algorithm, we construct a *general group-relative reinforcement learning gradient form* to create a **decoupled gradient estimator**, which is *strictly aligned* with Proposition 3.1. Its sole purpose is to eliminate the "structural asymmetric terms" identified in the proposition, without making commitments beyond the scope of the proposition regarding the indeterminacy of credit assignment under termination rewards.

4.1. Review of General Group-relative Gradient Structure

We begin by reviewing a general class of group-relative reinforcement learning objectives (linear segment, omitting clipping terms):

$$\mathcal{J}(\theta) = \mathbb{E}_{x,\, \{\tau_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} w_i(\tau_i; \theta) \, \widehat{A}_i \right], \tag{14}$$

where $\widehat{A}_i$ is the group-relative advantage, with zero mean constraint $\sum_{i=1}^{G} \widehat{A}_i = 0$, and $w_i(\tau_i;\theta)$ is the weight function dependent on the entire trajectory.

Using the logarithmic derivative trick, we obtain the following *unified gradient template*:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} w_i(\tau_i;\theta)\,\widehat{A}_i \;\cdot\; \sum_{t=1}^{T_i} \alpha_{i,t}(\theta)\,\mathbf{g}_{i,t}(\theta)\right],$$
$$\mathbf{g}_{i,t}(\theta) \triangleq \nabla_\theta \log \pi_\theta\big(a_t^{(i)} \mid x, a_{<t}^{(i)}\big) \tag{15}$$

where $\alpha_{i,t}(\theta) \geq 0$ is the coefficient induced by the weight aggregation form (e.g., length normalization corresponds to $\alpha_{i,t} = 1/T_i$).

4.2. Method Design Principle: Constructive Response to the Proposition

Based on the above proposition, our method follows an extremely conservative yet strictly aligned design principle:

> *We do not attempt to alter the token-level gradient direction, nor redefine the advantage or introduce additional supervision; we only eliminate the structural asymmetric terms introduced by sequence-level coupled weights that directly correspond to the proposition.*

To this end, we focus on the *trajectory-level coefficients*

$$w_i(\tau_i;\theta)\,\widehat{A}_i,$$

and apply a deterministic transformation only *within the group* to the weight vector $\mathbf{w} = (w_1, \ldots, w_G)$ to weaken or eliminate its ability to disrupt token-level gradient symmetry.

4.3. Decoupled Group-relative Gradient Estimator

Specifically, we define the following decoupled gradient estimator:

$$\widehat{\nabla_\theta J}_{\mathrm{dec}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \widetilde{w}_i\,\widehat{A}_i \cdot \sum_{t=1}^{T_i} \alpha_{i,t}\,\nabla_\theta \log \pi_\theta(a_t^{(i)} \mid x, a_{<t}^{(i)})\right] \tag{16}$$

where $\widetilde{\mathbf{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_G)$ is given by a *group transformation*

$$\widetilde{\mathbf{w}} = \mathcal{T}(\mathbf{w}, \widehat{\mathbf{A}})$$

and is implemented using a stop-gradient procedure.

4.4. Two Instantiations of Intra-group Transformations

We present two simple yet effective transformations, both aiming to suppress asymmetric random modulation introduced by sequence-coupled weights within the group, thereby recovering/approximating token-level cancellation structures.

4.4.1. Transformation 1: Group-Constant

Define

$$w_{\min} \triangleq \min_{j\in\{1,\ldots,G\}} w_j(\tau_j;\theta), \qquad \widetilde{w}_i \triangleq w_{\min},\ \ \forall i. \tag{17}$$

This transformation ensures that all trajectories within the group share the same weight scale, eliminating the non-canceling contributions from w_i differences within the group. It is important to note that this transformation does not zero the overall gradient, as gradients are a weighted sum of individual token gradients; the consistency in intra-group weights only guarantees cancellation recovery (or approximation) in the "shared token subspace where gradient directions align" (see Appendix D).

4.4.2. Transformation 2: Adv-Orthogonal Reweighting

The second transformation does not require equal $\widetilde{w}_i$ values but applies a *minimal disturbance reweighting* within the group to suppress systematic biases induced by sequence coupling. This is achieved by minimizing the correlation between weight vectors and advantage vectors in the inner product sense. The standard form is the orthogonal projection (in the stop-gradient sense):

$$\widetilde{\mathbf{w}} = \mathbf{w} - \frac{\widehat{\mathbf{A}}^\top \mathbf{w}}{\|\widehat{\mathbf{A}}\|_2^2}\,\widehat{\mathbf{A}}. \tag{18}$$

If non-negativity constraints are required, the Positive Orth-Proj/QP or its closed-form approximation in Appendix B can be used.

Alignment with the Proposition (Why it Mitigates Learning Tax and Entropy Collapse). Proposition 3.1 indicates that the common structural source of learning tax and entropy collapse is the token-level asymmetric and non-canceling gradients caused by sequence-coupled weights. Both transformations weaken or eliminate intra-group weight differences (or their correlation with advantages), recovering (or approximating) the intra-group cancellation structure in the shared token subspace, thus systematically reducing ineffective updates to weakly rewarded tokens and suppressing probabilistic drift between equivalent correct solutions.

4.5. Testable Predictions: Reduced Learning Tax ⇒ Better Endpoint Performance and More Stable Convergence

This section presents *testable predictions* directly derived from the structural fixes in Section 4, along with verification methods on HMMT25, AIME25, and LiveCodeBench.

Prediction 1 (Computational Efficiency). If intra-group transformations effectively suppress ineffective updates and exclude interference, learning efficiency may improve, leading to faster achievement of fixed performance thresholds under compute-matched settings:

$$\mathrm{Steps}\big(\mathrm{Score} \geq \kappa\big) \downarrow, \tag{19}$$

where Score is the evaluation metric for the corresponding benchmark.

Prediction 2 (Convergence Stability). If intra-group transformations focus the effective gradient, local oscillations in the training curve should decrease. We measure this with the second-order difference jitter metric:

$$\mathrm{Jitter}_2(m) \triangleq \frac{1}{T-2} \sum_{t=1}^{T-2} \big|m_{t+2} - 2m_{t+1} + m_t\big|, \tag{20}$$

and predict under compute-matched conditions:

$$\mathrm{Jitter}_2\big(m^{\mathrm{ours}}\big) \; < \; \mathrm{Jitter}_2\big(m^{\mathrm{base}}\big), \tag{21}$$

Prediction 3 (Endpoint Performance). Ineffective updates repeatedly impose biases on high-frequency tokens/templates unrelated to the reward, equivalent to continuous reshaping of the surface distribution (distributional drift), causing parameter drift and performance regression on unoptimized tasks/subdistributions (catastrophic forgetting). Reducing these ineffective updates could lead to higher endpoint performance:

$$\mathrm{Score}_{\mathrm{final}} \uparrow . \tag{22}$$

5. Experiments

This section validates the three predictions from Section 4.5 under the compute-matched protocol. We compare the baseline methods with our approach (two types of intra-group transformations: Min-Replace / Orth-Proj). To ensure fairness, all methods are matched with the same training compute budget (same total token generation and parameter update steps), with identical training and evaluation settings. Details of the DFPO (Drift Fixing Policy Optimization, implemented with our intra-group transformation) algorithm and hyperparameters can be found in Appendix B.

Tasks and Datasets. We evaluate the proposed method on a set of mathematical and code reasoning benchmarks, including *HMMT25* (Balunović et al., 2025), *AIME25* (Mathematical Association of America, 2025), *LiveCodeBench v6 (25.02-25.05)* (Jain et al., 2024).

Base Model. Qwen3-32B (Team, 2025), Qwen3-Next-80B-A3B-Thinking (Yang et al., 2025a).

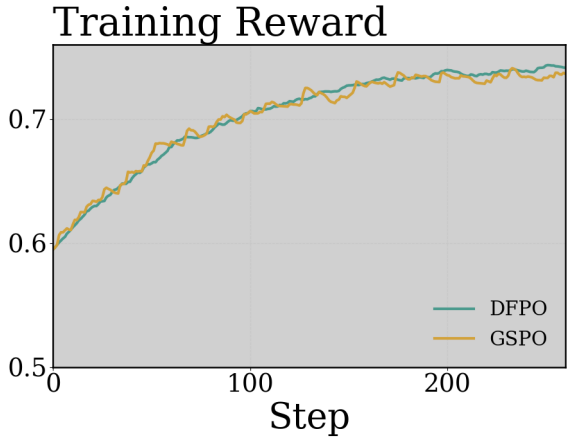


Figure 1. Training curves on Qwen3-Next-80B-A3B-Thinking show that under **compute-matched** settings, **DFPO** achieves substantially higher training efficiency than GSPO.

Baseline Methods and Comparison Settings. (1) GSPO; (2) GRPO; (3) GRPO-fix, which fixes the asymmetric pruning in GRPO based on our design principles; algorithm details are in Appendix H. Experimental parameter configurations are provided in Appendix I.

Compute-Matched Protocol. We match the total training compute by satisfying the following constraints: (i) the total number of generated tokens is the same, (ii) the optimizer parameter update steps are the same, (iii) each prompt has the same rollout budget. All methods use identical models, decoding strategies, batch schemes, and hardware configurations.

6. Results and Analysis

As shown in Figure 2 and Table 2, in the **compute-matched** setting, DFPO reaches a fixed training reward threshold with less computation, validating Prediction 1. Additionally, DFPO's training return curve is smoother with fewer drawdowns. Short-term oscillations are quantitatively captured by the second-order difference jitter $\mathrm{Jitter}_2(\cdot)$ (Equation (20)), satisfying $\mathrm{Jitter}_2(m^{\mathrm{DFPO}}) < \mathrm{Jitter}_2(m^{\mathrm{GSPO}})$ (Equation (21)), validating Prediction 2. Furthermore, Table 1 shows that DFPO achieves higher endpoint performance on AIME25, LiveCodeBench, and HMMT25, validating Prediction 3.

6.1. Mechanism Verification: Sequence Coupling Disruption and Decoupling Restoration

This section reports two mechanism metrics: (i) the asymmetry of intra-group gradient modulation and (ii) the proportion of ineffective updates for high-frequency tokens.

Method	Qwen3-32B Acc avg@32 (%)			Qwen3-Next Acc avg@32 (%)		
	AIME25	LiveCodeBench	HMMT25	AIME25	LiveCodeBench	HMMT25
DFPO (Min-Replace)	82.5±1.1	71.6±0.7	**61.4**±1.5	**93.2**±0.9	75.1±0.8	80.1±1.1
DFPO (Orth-Proj)	**82.6**	**71.6**	61.3	93.1	**75.2**	**80.2**
GSPO	76.9±1.3	64.7±1.5	55.8±0.9	89.8±1.7	71.0±1.4	75.8±1.2
GRPO	76.9±1.2	64.5±0.8	55.5±1.5	89.7±1.2	70.9±0.9	75.5±1.8
GRPO-fix	80.6±1.4	69.1±1.2	59.6±0.9	91.9±1.5	73.9±1.1	79.4±0.7

Table 1. Results on Qwen3-32B and Qwen3-Next-80B-A3B-Thinking. We report the mean and its 95% bootstrap confidence interval (mean ± 95% CI) over 5 random seeds; the improvement over baseline methods is statistically significant under paired bootstrap test (p ¡ 0.01). Inference settings are provided in Appendix J.

Method	Compute@Reward=0.70
GSPO	1.00×
DFPO	**0.91**×

Table 2. On Qwen3-Next-80B-A3B-Thinking, the compute required to reach a fixed training reward threshold under a matched compute budget. Results are normalized relative to GSPO.

Asymmetry of Gradient Modulation.

$$\mathrm{Asym}(t) = \mathrm{Var}_{i\in\{1,\dots,G\}}\left(w_i(\tau_i;\theta)\,\widehat{A}_i\right), \qquad (23)$$

where $w_i\widehat{A}_i$ is the trajectory-level modulation coefficient. Larger Asym indicates more difficulty in forming intra-group cancellation for shared/similar tokens. Decoupling transformations significantly reduce Asym in experiments.

Energy on Frequent Tokens.

$$\mathrm{Energy}(B) = \frac{\sum_{t\in B}\|\nabla_\theta \ell_t\|_2}{\sum_t \|\nabla_\theta \ell_t\|_2}, \qquad (24)$$

where B is the token set bucketed by frequency. Compared to the baseline, our method reduces Energy for high-frequency buckets, indicating a decrease in reward-irrelevant updates (learning tax).

6.2. Ablation Study

Model Variant	AIME25 (%)	LiveCodeBench (%)	HMMT25 (%)
DFPO(Min-Replace)	**82.5**	**71.6**	**61.4**
GSPO (baseline)	76.9	64.7	55.8
DFPO (no stop-grad)	75.6	63.8	54.9
DFPO(scale by 0.5)	75.1	63.2	54.1

Table 3. Qwen3-32B.Ablation results.

DFPO (no stop-grad): Removing the stop-grad Min-Replace method.

DFPO (scale by 0.5): Replacing equation 50 with: $\widetilde{s}_i(\theta) \triangleq 0.5\,\bar{s}_i(\theta)$.

Table 3 presents an ablation study on two key design choices of DFPO(Min-Replace). First, removing the stop-gradient (DFPO no stop-grad) consistently degrades performance on AIME25, LiveCodeBench, and HMMT25, indicating that including groupwise transformations in the backward pass introduces additional gradient coupling and instability. Freezing gradients as a "control variable" within the group is thus essential for method stability. Second, to rule out the explanation that performance gains are merely due to smaller updates, we replace groupwise consistency with simple global scaling (DFPO scale by 0.5), which results in significantly worse performance compared to the full Min-Replace. This shows that the performance boost comes not from a more conservative step size, but from groupwise consistency's structural correction of sequence coupling weights, weakening asymmetric modulation within the group and recovering (or approximating) the required cancellation structure in the shared/high-frequency token subspace. Overall, these ablations align with the mechanism analysis in this paper regarding the necessary conditions for exchangeability-cancellation.

Group size G and relative gains. Table 4 reports the relative improvement Δ (DFPO Min-Replace−GSPO). Increasing G leads to more heterogeneous trajectories within the group, making the baseline more susceptible to non-offset shared/high-frequency tokens and the accumulation of learning tax. DFPO's intra-group alignment recovers or approximates offset within this subspace, yielding a greater relative gain.

7. Conclusion

This paper formalizes a recurrent training instability into a *structural boundary*: under sparse termination rewards, the stability of group-based reinforcement learning is constrained by the *token-wise exchangeability* of the objective function. We show through gradient decomposition and minimal counterexamples that sequence-level multiplicative coupling breaks this symmetry, leading to gradient cancellation failure and inducing learning tax accumulation

and entropy collapse. We argue that restoring this symmetry is a *necessary but not sufficient* condition to mitigate these issues. Based on this, we propose two minimal intra-group transformations to restore the cancellation structure on shared tokens. Experimental results validate our testable predictions, confirming the explanatory and practical value of the structural boundary.

8. Limitations

(1) This paper characterizes the structural *necessary condition*; with only termination rewards, credit assignment remains unidentifiable, so this fix can only alleviate or delay instability, not guarantee its complete removal. (2) The key derivation exposes the structural difference between "decomposable vs. coupled"; its interaction with mechanisms like pruning and normalization requires further analysis. (3) Projection-based transformations may introduce biases. We use them as the minimal viable construct to verify the proposition, without fully exploring the optimal implementation space or broader baseline coverage.

Ethical Considerations

All experiments are based on publicly available benchmark datasets, with no involvement of personal or sensitive information. Thus, within the current experimental setup and research scope, no apparent ethical risks are present.

Reproducibility Statement

To ensure reproducibility, we provide detailed information on all code, datasets, and models used. The relevant implementations will be shared anonymously via public repositories like GitHub, enabling other researchers to replicate the experiments.

References

Balunović, M., Dekoninck, J., Petrov, I., Jovanović, N., and Vechev, M. Matharena: Evaluating llms on uncontaminated math competitions, February 2025. URL https://matharena.ai/.

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Mathematical Association of America. 2025 AIME I and AIME II Problems and Solutions, 2025. URL https://artofproblemsolving.com/wiki/index.php/2025_AIME_I_Problems. Accessed: Jan 6, 2026.

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

Team, Q. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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

Yang, K., Wang, Y., Li, Z., et al. Sspo: Subsentence-level policy optimization. *arXiv preprint arXiv:2511.04256*, 2025b.

Yang, S., Dou, C., Guo, P., Lu, K., Ju, Q., Deng, F., and Xin, R. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*, 2025c.

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

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

G	2	4	8	16
avg@32	3.1	4.2	5.6	5.7

Table 4. Group size ablation (DFPO Min-Replace score minus GSPO score; Qwen3-32B; AIME25).

A. Unified Toy Problem: Learning Tax and Entropy Collapse

We present a unified toy problem using GSPO as an example, illustrating two core failure modes identified in Proposition 3.1: (i) learning tax caused by ineffective updates to tokens unrelated to the reward, and (ii) entropy collapse between semantically equivalent correct solutions. The example is intentionally minimal, fully algebraic, and independent of any language-specific surface forms.

A.1. Prompt and Trajectories

Consider the prompt:

$$x : \text{"What is 10+10?"}$$

and the following three trajectories:

$$\tau_1 = (\text{"The answer is 25."}) \qquad \text{(incorrect)}, \tag{25}$$
$$\tau_2 = (\text{"The answer is 20."}) \qquad \text{(correct)}, \tag{26}$$
$$\tau_3 = (\text{"10+10=20."}) \qquad \text{(correct)}. \tag{27}$$

The reward function depends only on the semantic correctness of the final answer:

$$R(\tau_1) = 0, \quad R(\tau_2) = R(\tau_3) = 1. \tag{28}$$

We consider a group of size $G = 3$ containing the above three trajectories. The group relative advantage is defined as:

$$\widehat{A}_i = R(\tau_i) - \frac{1}{3}\sum_{j=1}^{3} R(\tau_j), \tag{29}$$

resulting in:

$$\widehat{A}_1 = -\frac{2}{3}, \quad \widehat{A}_2 = \widehat{A}_3 = \frac{1}{3}. \tag{30}$$

Let $r_{i,t}(\theta)$ be the token-level importance ratio at time step t, and define the sequence-level weight as:

$$s_i(\theta) = \prod_t r_{i,t}(\theta). \tag{31}$$

A.2. Learning Tax: Ineffective Updates to Reward-Unrelated Tokens

Minimal Case ($G = 2, T = 3$): Sequence coupling causes non-canceling gradient for shared prefixes. To highlight the structural origin, we first consider a minimal group containing two trajectories $\{\tau_1, \tau_2\}$, with their group relative advantage satisfying the zero-mean constraint:

$$\widehat{A}_2 = -\widehat{A}_1 \triangleq A, \quad A > 0. \tag{32}$$

The two trajectories are written as token sequences of length $T = 3$:

$$\tau_1 = (a_1 = \text{"answer"}, a_2 = \text{"is"}, a_3 = \text{"25"}), \tag{33}$$
$$\tau_2 = (a_1 = \text{"answer"}, a_2 = \text{"is"}, a_3 = \text{"20"}). \tag{34}$$

Assuming they have the same context-token pairs in the first two steps:

$$r_{1,1} = r_{2,1} = \rho_1, \quad r_{1,2} = r_{2,2} = \rho_2, \tag{35}$$

but differing tokens in the last step:

$$r_{1,3} = \lambda_1, \quad r_{2,3} = \lambda_2, \quad \lambda_1 \neq \lambda_2. \tag{36}$$

Under sequence coupling weighting:

$$s_1 = \rho_1\rho_2\lambda_1, \quad s_2 = \rho_1\rho_2\lambda_2. \tag{37}$$

Thus, for any shared prefix token ($t = 1$ or $t = 2$), the effective gradient coefficient along $\nabla_\theta \log \pi_\theta(a_t \mid h_t)$ is proportional to:

$$\widehat{A}_1 s_1 + \widehat{A}_2 s_2 = (-A)\rho_1\rho_2\lambda_1 + A\rho_1\rho_2\lambda_2 = A\rho_1\rho_2(\lambda_2 - \lambda_1), \tag{38}$$

which is strictly non-zero when $\lambda_2 \neq \lambda_1$. This shows that *even if the reward depends only on the final token, sequence coupling importance weighting still prevents the cancellation of shared prefix gradients within the group*, leading to systematic updates on reward-unrelated tokens, the minimal counterexample of learning tax.

A.3. Entropy Collapse: Probability Drift Between Equivalent Correct Solutions

Although τ_2 and τ_3 differ in surface form, they express the same mathematical fact and are equally correct; they also have the same group relative advantage: $\widehat{A}_2 = \widehat{A}_3 = \frac{1}{3}$. However, due to differences in tokenization, length, and local likelihood, their sequence-level weights generally differ:

$$s_2(\theta) \neq s_3(\theta). \tag{39}$$

In the linear approximation region of updates:

$$\Delta \log \pi_\theta(\tau_i) \approx \eta \cdot c \cdot \widehat{A}_i s_i(\theta), \quad c > 0, \tag{40}$$

the logarithmic probability ratio of the two correct trajectories is:

$$\log \frac{\pi_{\theta^+}(\tau_2)}{\pi_{\theta^+}(\tau_3)} \approx \log \frac{\pi_\theta(\tau_2)}{\pi_\theta(\tau_3)} + \eta c \left(\widehat{A}_2 s_2(\theta) - \widehat{A}_3 s_3(\theta) \right). \tag{41}$$

Even if $\widehat{A}_2 = \widehat{A}_3$, any small difference between $s_2(\theta)$ and $s_3(\theta)$ will cause the probability ratio to drift; this drift accumulates multiplicatively through repeated updates, concentrating probability mass on one surface form, leading to a reduction in policy entropy over the semantically equivalent correct answer space, manifesting as entropy collapse.

A.4. Discussion

This unified toy problem demonstrates that both learning tax and entropy collapse stem from the same structural origin: sequence coupling importance weighting disrupts token-level gradient symmetry. This example is highly concise yet captures key failure modes observed in large-scale training; and these phenomena occur independently of specific languages, reward designs, or implementation details, highlighting their structural nature.

B. Implementation Details: An In-Group Reweighting Instance of a Sequence-Coupled Objective

The goal of this section is not to propose an engineering improvement to a specific baseline, but rather **to validate the structural proposition in this paper about "sequence-coupled weights breaking token-level gradient symmetry"**. We choose a representative group-relative reinforcement learning algorithm as the *vehicle* for instantiating this proposition. Specifically, we **arbitrarily** adopt GSPO as the analysis and implementation platform, as its objective function contains both: (i) group-relative advantages and (ii) sequence-level (coupled) importance weighting, thus directly supporting the structural phenomena outlined in Section 3.

Without altering GSPO's *group-relative advantage estimation* and *sequence-level clipping* framework, we apply a *minimized in-group transformation* to the sequence-level importance ratio vectors within each group, resulting in DFPO (Drift Fixing Policy Optimization). The sole purpose of this transformation is to enforce the critical orthogonality condition $\sum_i \widetilde{s}_i \widehat{A}_i = 0$ within the group, thus restoring the gradient cancellation structure on shared tokens and transforming the "structural sources of learning tax/entropy collapse" into empirically verifiable differences.

B.1. GSPO Formula

GSPO adopts the following sequence-level optimization objective:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(s_i(\theta)\widehat{A}_i,\ \text{clip}\left(s_i(\theta),1-\varepsilon,1+\varepsilon\right)\widehat{A}_i\right)\right], \tag{42}$$

where the group-relative advantage is estimated as:

$$\widehat{A}_i = \frac{r(x,y_i) - \text{mean}\left(\{r(x,y_i)\}_{i=1}^{G}\right)}{\text{std}\left(\{r(x,y_i)\}_{i=1}^{G}\right)}, \tag{43}$$

and the sequence-level importance ratio is defined based on sequence likelihood normalization:

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)}\right)^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}\right). \tag{44}$$

By definition, the group-relative advantages satisfy zero mean: $\sum_{i=1}^{G}\widehat{A}_i = 0$.

B.2. DFPO: In-Group Transformation on Sequence Weights

Let the in-group sequence weight vector and advantage vector be

$$\mathbf{s}(\theta) \triangleq (s_1(\theta),\ldots,s_G(\theta))^\top, \qquad \widehat{\mathbf{A}} \triangleq (\widehat{A}_1,\ldots,\widehat{A}_G)^\top. \tag{45}$$

The core of DFPO is: applying a transformation $\mathbf{s}(\theta) \mapsto \widetilde{\mathbf{s}}(\theta)$ within each group, and replacing $s_i(\theta)$ in the GSPO objective with $\widetilde{s}_i(\theta)$.

Due to clipping, we need to first extract $\widehat{A}$, clip it, and then apply the in-group transformation. However, note that:

$$\min(s_i\widehat{A}_i,\ \text{clip}(s_i)\widehat{A}_i) \neq \widehat{A}_i\min(s_i,\ \text{clip}(s_i)) \quad \text{when } \widehat{A}_i < 0 \text{ (this does not hold).}$$

This is because a negative $\widehat{A}_i$ reverses the inequality, so a *sign-consistent* rewriting is required.

B.3. Step 1: Rewriting GSPO as "Clipping First, Then Multiplying by Advantage"

Define

$$c_i(\theta) \triangleq \text{clip}(s_i(\theta),\,1-\varepsilon,\,1+\varepsilon), \qquad s_i(\theta) = \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)}\right)^{\frac{1}{|y_i|}}. \tag{46}$$

We introduce a **"sign-aware" post-clip weight**:

$$\bar{s}_i(\theta) \triangleq \begin{cases}\min\big(s_i(\theta),\,c_i(\theta)\big), & \widehat{A}_i \geq 0,\\ \max\big(s_i(\theta),\,c_i(\theta)\big), & \widehat{A}_i < 0.\end{cases} \tag{47}$$

Thus, the original GSPO objective

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\big(s_i(\theta)\widehat{A}_i,\ c_i(\theta)\widehat{A}_i\big)\right] \tag{48}$$

can be strictly equivalently rewritten as

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\widehat{A}_i\,\bar{s}_i(\theta)\right]. \tag{49}$$

Equivalence Explanation (One Line). When $\widehat{A}_i \geq 0$, we have $\min(s_i\widehat{A}_i, c_i\widehat{A}_i) = \widehat{A}_i \min(s_i, c_i)$; when $\widehat{A}_i < 0$, we have $\min(s_i\widehat{A}_i, c_i\widehat{A}_i) = \widehat{A}_i \max(s_i, c_i)$, thus yielding (49).

B.4. Step 2: In-Group Transformation on "Post-Clipped Weight Vector" (DFPO)

The following two transformations correspond to the previously defined Min-Replace and Orth-Proj, but note that they act on $\bar{\mathbf{s}}$ (the "post-clipped weights").

B.4.1. Transformation 1: Min-Replace (Consistency on Post-Clipped Weights)

$$\bar{s}_{\min}(\theta) \triangleq \min_{j\in\{1,\dots,G\}} \bar{s}_j(\theta), \qquad \widetilde{s}_i(\theta) \triangleq \bar{s}_{\min}(\theta) \ \ \forall i. \tag{50}$$

B.4.2. Transformation 2: Positive Orth-Proj (Non-Negative Constrained Orth-Proj)

In practice, the weights must be non-negative, and using the orthogonal projection in (63) could result in negative components. Therefore, we replace the "orthogonality constraint"

$$\widehat{\mathbf{A}}^\top \widetilde{\mathbf{s}} = 0$$

with a **minimally disturbed projection in the non-negative domain**: maintaining $\widetilde{\mathbf{s}} \geq \mathbf{0}$ while making it as close as possible to the original post-clipped weights $\bar{\mathbf{s}}$, and satisfying group orthogonality.

Specifically, we define the **Positive Orth-Proj** as the following quadratic program (QP):

$$\widetilde{\mathbf{s}}(\theta) = \arg\min_{\mathbf{v}\in\mathbb{R}^G} \ \frac{1}{2}\|\mathbf{v} - \bar{\mathbf{s}}(\theta)\|_2^2 \quad \text{s.t.} \quad \widehat{\mathbf{A}}^\top \mathbf{v} = 0, \ \ \mathbf{v} \succeq \mathbf{0}. \tag{51}$$

Feasibility and Intuitive Explanation. Since the group-relative advantages satisfy $\sum_{i=1}^G \widehat{A}_i = 0$, the hyperplane constraint $\widehat{\mathbf{A}}^\top \mathbf{v} = 0$ typically has a non-trivial feasible solution in the positive orthogonal cone, especially when the advantage vector contains both positive and negative components (where mass can be redistributed between the positive/negative advantage subsets to achieve a zero inner product). In practice, (51) can be solved by a standard QP solver or approximated using an efficient projection algorithm.

Simplified Implementation: Truncate-and-Rebalance. We use the following **two-step closed-form approximation** to ensure non-negativity while restoring orthogonality in a single correction step:

Step 1 (Truncate to Non-Negative).

$$\mathbf{s}^+(\theta) = \max\big(\bar{\mathbf{s}}(\theta), \epsilon\mathbf{1}\big) \quad \text{(element-wise)}. \tag{52}$$

Step 2 (Minimal Rebalancing in the Non-Negative Domain). Let the positive/negative advantage index sets be $\mathcal{P} \triangleq \{i : \widehat{A}_i > 0\}, \mathcal{N} \triangleq \{i : \widehat{A}_i < 0\}$. Calculate the current inner product deviation

$$\delta(\theta) \triangleq \widehat{\mathbf{A}}^\top \mathbf{s}^+(\theta) = \sum_{i=1}^G \widehat{A}_i s_i^+(\theta). \tag{53}$$

We only scale on one "side" proportionally to bring the inner product back to zero while maintaining non-negativity:

$$\widetilde{\mathbf{s}}(\theta) = \begin{cases} \Big(\alpha\, \mathbf{s}^+_{\mathcal{P}}(\theta), \ \mathbf{s}^+_{\mathcal{N}}(\theta) \Big), & \delta(\theta) > 0, \\ \Big(\mathbf{s}^+_{\mathcal{P}}(\theta), \ \beta\, \mathbf{s}^+_{\mathcal{N}}(\theta) \Big), & \delta(\theta) < 0, \\ \mathbf{s}^+(\theta), & \delta(\theta) = 0, \end{cases} \tag{54}$$

where the scaling coefficients are

$$\alpha = \frac{-\sum_{i\in\mathcal{N}} \widehat{A}_i s_i^+}{\sum_{i\in\mathcal{P}} \widehat{A}_i s_i^+} \in (0, 1], \qquad \beta = \frac{-\sum_{i\in\mathcal{P}} \widehat{A}_i s_i^+}{\sum_{i\in\mathcal{N}} (-\widehat{A}_i) s_i^+} \in (0, 1]. \tag{55}$$

By construction, $\widetilde{\mathbf{s}}(\theta) \succeq \epsilon\mathbf{1}$ and strictly satisfies $\widehat{\mathbf{A}}^\top \widetilde{\mathbf{s}}(\theta) = 0$.

Discussion. (51) provides a strict definition of "non-negative orthogonal projection"; (52)–(55) offer an engineering-friendly version that does not require iterative re-projection: it only performs a single truncation and one side scaling to precisely restore the group orthogonality constraint while maintaining non-negativity.

C. Bias of Min-Replace and the "No Reverse Update" Property

This appendix clarifies an important, often confused point in this paper:

Does applying Min-Replace (e.g., taking the minimum value within a group) to group weights break the unbiasedness of importance sampling (IS), and does it cause "wrong-direction" reverse updates? This section rigorously explains: Although Min-Replace may introduce bias, it *does not cause reverse updates* in the direction of the policy gradient update we care about. Its main effect is a conservative scaling of the update magnitudes for each trajectory, which only slows down convergence (with a smaller effective step size).

Linear Region. We analyze in a local linear region (ignoring the piecewise points of $\min$ /clip; this assumption is made solely to avoid piecewise discussion and does not affect the directional conclusions). For any trajectory $\tau_i = (a_1^{(i)}, \ldots, a_{T_i}^{(i)})$ within a group, define the length-normalized sequence-level ratio (consistent with Eq. (44)) as

$$s_i(\theta) = \left(\frac{\pi_\theta(\tau_i \mid x)}{\pi_{\theta_{\text{old}}}(\tau_i \mid x)}\right)^{\frac{1}{T_i}} = \exp\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\log r_{i,t}(\theta)\right), \qquad r_{i,t}(\theta) = \frac{\pi_\theta(a_t^{(i)} \mid h_t^{(i)})}{\pi_{\theta_{\text{old}}}(a_t^{(i)} \mid h_t^{(i)})}. \tag{56}$$

Note that $s_i(\theta) > 0$ always holds.

Direction of the Original (Sequence-Coupled) Gradient Term. In the linear region, a typical sequence-coupled group relative objective can be written as

$$\mathcal{J}_{\text{seq}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\widehat{A}_i\, s_i(\theta). \tag{57}$$

For any trajectory i and token t, we have

$$\nabla_\theta s_i(\theta) = s_i(\theta)\cdot\frac{1}{T_i}\sum_{t'=1}^{T_i}\nabla_\theta \log \pi_\theta(a_{t'}^{(i)} \mid h_{t'}^{(i)}). \tag{58}$$

Therefore, the contribution direction for any token is aligned with $\nabla_\theta \log \pi_\theta(\cdot)$, modulated by the coefficient $\widehat{A}_i s_i(\theta)$. Since $s_i(\theta) > 0$, the update sign for each token in trajectory i is determined by $\widehat{A}_i$.

Min-Replace: Conservative Proportional Scaling for Sequence Weights. Consider applying Min-Replace (or more generally, "conservative scaling") to the weights within each group:

$$\widetilde{s}_i(\theta) = \phi_i(\mathbf{s}(\theta))\cdot s_i(\theta), \qquad \mathbf{s}(\theta) = (s_1(\theta), \ldots, s_G(\theta)), \tag{59}$$

where the scaling factor satisfies

$$0 < \phi_i(\mathbf{s}(\theta)) \le 1. \tag{60}$$

Min-Replace (taking the group minimum and applying it for contraction) is an extreme case of this class: For "non-minimal" trajectories, ϕ_i is significantly less than 1, while for the minimal trajectory, $\phi_i = 1$.

By replacing s_i with $\widetilde{s}_i$ in Eq. (57), we obtain the corrected objective

$$\widetilde{\mathcal{J}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\widehat{A}_i\,\widetilde{s}_i(\theta) = \frac{1}{G}\sum_{i=1}^{G}\widehat{A}_i\,\phi_i(\mathbf{s}(\theta))\,s_i(\theta). \tag{61}$$

Theorem C.1 (No Reverse Update: Min-Replace Only Causes Proportional Scaling)**.** *Under the assumption that Eq. (60) holds and the gradient terms of ϕ_i are neglected (see discussion below), the corrected gradient direction is* aligned *with*

the original gradient direction for each trajectory and token: For any i, t, the coefficient for $\nabla_\theta \log \pi_\theta(a_t^{(i)} \mid h_t^{(i)})$ in the corrected update satisfies

$$\underbrace{\widehat{A}_i\, \widetilde{s}_i(\theta)}_{\textit{Min-Replace coefficient}} = \phi_i(\mathbf{s}(\theta)) \quad \underbrace{\widehat{A}_i\, s_i(\theta)}_{\textit{Original coefficient}}, \qquad \phi_i(\mathbf{s}(\theta)) \in (0, 1]. \tag{62}$$

Therefore, Min-Replace does not flip the update direction (no negative scaling), and its only effect is to shrink the update magnitudes, equivalent to a smaller effective learning rate, which may slow down but stabilize convergence.

Proof. From Eq. (59), we directly get $\widehat{A}_i\widetilde{s}_i = \widehat{A}_i\phi_i s_i$. Since $s_i > 0$ and $\phi_i \in (0, 1]$, $\widehat{A}_i\widetilde{s}_i$ and $\widehat{A}_i s_i$ have the same sign. Thus, for any token's score function term, the corrected gradient is a positive proportional scaling of the original gradient, with no sign reversal (no reverse update). This concludes the proof. □

Precise Definition and Impact of "Loss of IS Unbiasedness." Importance sampling unbiasedness typically requires the weights to satisfy a Radon-Nikodym derivative form for some target distribution. Min-Replace clearly no longer satisfies this strict condition and may introduce bias. However, the above theorem shows that this bias does not manifest as "reducing the probability of tokens that should increase," but instead scales down the updates for all tokens on each trajectory in a proportional manner (more conservatively). In the token one-shot setting, this scaling does not introduce additional sign inconsistencies within the trajectory, and primarily results in reduced variance/smaller effective step size (slower but more stable training).

D. Why $\sum_{i=1}^{G} \widetilde{s}_i \widehat{A}_i = 0$ Does Not Imply Zero Gradient

This appendix clarifies a common confusion that is crucial in this paper: Even if we construct weights $\widetilde{s}_i$ within each group such that

$$\sum_{i=1}^{G} \widetilde{s}_i\, \widehat{A}_i = 0, \tag{63}$$

it *does not imply* that the corresponding policy gradient update is zero. Intuitively, (63) only constrains the "weighted sum of the group" to be zero, whereas the policy gradient is a weighted sum of "scalar weights × gradient vectors of each token in each trajectory"; unless these gradient vectors are identical across the group, a zero scalar does not necessarily make the vector sum zero.

Key Implementation Convention: Do Not Backpropagate $\widetilde{s}_i$ (Stop-gradient). In practical algorithms, $\widetilde{s}_i$ is derived from the sequence weight vector $\mathbf{s}(\theta)$ of the current group samples $(\{y_i\}, \{\widehat{A}_i\})$ through an in-group transformation. To ensure this transformation serves as a "structural correction/control variable," we adopt the *stop-gradient* convention: we treat $\widetilde{s}_i$ as a constant during backpropagation and do not differentiate the transformation operator itself. This convention aligns with the common handling in algorithms like PPO/GRPO/GSPO, where "sampling from $\pi_{\theta_{\text{old}}}$ and stopping gradients for advantage $\widehat{A}_i$" is applied.

Under this convention, (63) does not cause the overall gradient to degenerate to zero; instead, it restores the cancellation structure in the "shared token subspace of gradient direction" (as detailed in the following derivation).

D.1. Gradient Review for GSPO (Excluding Clipping Terms)

For comparison, we start with GSPO. The objective (linear segment) of GSPO is:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, \{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} s_i(\theta) \widehat{A}_i \right], \tag{64}$$

where

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)} \right)^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})} \right). \tag{65}$$

Using the log-derivative trick, the gradient is:

$$\nabla_\theta \mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} s_i(\theta)\widehat{A}_i \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right]. \tag{66}$$

In contrast, the GRPO (token-factorized) gradient is:

$$\nabla_\theta \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right]. \tag{67}$$

D.2. Gradient with In-Group Transformation: Why Not Zero

Now consider applying the in-group transformation $\mathbf{s} \mapsto \widetilde{\mathbf{s}}$ to the sequence weights and applying *stop-gradient* to $\widetilde{\mathbf{s}}$ during the update. That is, treat $\widetilde{s}_i$ as a constant coefficient during gradient calculation, without differentiating it.

The "DFPO-like" GSPO gradient is then:

$$\nabla_\theta \widetilde{\mathcal{J}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \widetilde{s}_i\, \widehat{A}_i \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right], \tag{68}$$

where $\widetilde{s}_i$ is not backpropagated with respect to θ (treated as a constant coefficient).

Note that (68) is a weighted sum of *vectors*:

$$\sum_i \widetilde{s}_i \widehat{A}_i \Big(\frac{1}{|y_i|}\sum_t \nabla_\theta \log \pi_\theta(\cdot)\Big).$$

Even if the scalar constraint $\sum_i \widetilde{s}_i \widehat{A}_i = 0$ is satisfied, it does not imply that the vector sum is zero unless the gradient vectors inside the parentheses are identical across the group. This is the key to understanding why "(63) restores cancellation for shared tokens, but does not erase the overall learning signal."

D.3. Formal Decomposition: Which Terms Get Cancelled and Which Don't

Let the gradient direction of a single token be

$$\mathbf{g}_{i,t}(\theta) \triangleq \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}). \tag{69}$$

Expanding (68), we get:

$$\nabla_\theta \widetilde{\mathcal{J}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|} \underbrace{\left(\frac{\widetilde{s}_i \widehat{A}_i}{|y_i|}\right)}_{\text{scalar coefficient}} \mathbf{g}_{i,t}(\theta)\right]. \tag{70}$$

Strict Cancellation in the Shared Token Subset. Consider a "shared context-token" event: There exists a fixed context-action pair $(h^\star, a^\star)$ such that certain trajectories in the group have matching tokens at specific time steps:

$$(y_{i,<t}, y_{i,t}) = (h^\star, a^\star),$$

leading to identical gradient directions:

$$\mathbf{g}_{i,t}(\theta) = \mathbf{g}^\star(\theta) \quad \text{for all } (i,t) \text{ in the shared event.} \tag{71}$$

Pulling out these shared terms from (70) results in their combined contribution:

$$\mathbf{g}^\star(\theta) \cdot \left(\sum_{(i,t)\in\mathcal{S}(h^\star,a^\star)} \frac{\widetilde{s}_i \widehat{A}_i}{|y_i|}\right), \tag{72}$$

where $\mathcal{S}(h^\star, a^\star)$ denotes the set of indices for the shared context-token pair.

If the length normalization factors are consistent (or nearly so) within this shared event, such as $|y_i|$ being the same or differences negligible, the term in parentheses in (72) is approximately proportional to $\sum_{i\in\mathcal{I}(h^\star,a^\star)} \widetilde{s}_i \widehat{A}_i$. Thus, when we construct $\widetilde{\mathbf{s}}$ within the group to satisfy

$$\sum_{i=1}^{G} \widetilde{s}_i \widehat{A}_i = 0, \tag{73}$$

we can suppress/cancel the combined gradient for the "shared token subspace." This forms the basis for our in-group transformation mechanism to restore token-level symmetry (exchangeability).

Gradients on Non-shared Tokens Generally Not Zero. However, for general tokens, $\mathbf{g}_{i,t}(\theta)$ differs across trajectories and time steps, so (70) is a weighted sum of *vectors* in different directions. Even if the scalar constraint (63) holds, it does not imply that these different direction vectors will fully cancel. Formally,

$$\sum_{i=1}^{G} \widetilde{s}_i \widehat{A}_i = 0 \;\not\Rightarrow\; \sum_{i,t} \frac{\widetilde{s}_i \widehat{A}_i}{|y_i|} \mathbf{g}_{i,t}(\theta) = \mathbf{0}, \tag{74}$$

unless all $\mathbf{g}_{i,t}(\theta)$ are collinear and the coefficients strictly match, which is an extremely rare degenerate case.

Thus, the role of (63) is not to "make the update zero," but to *selectively* eliminate/suppress the systematic drift of tokens that repeat within the group, are weakly correlated with rewards, and have highly consistent gradient directions. This reduces learning tax and delays entropy collapse, while not blocking the learning signal for truly reward-related decision tokens.

E. Learning Tax: Cumulative Effects of Ineffective Updates and Their Severe Consequences

Background and Definition (For Appendix Consistency). In sequence-based reinforcement learning fine-tuning with terminal rewards (e.g., group-relative methods and their variants), we define *ineffective updates* or *learning tax* as: parameters being continuously updated, but these updates statistically **do not bring net gains in the desired capability**, and may even be unrelated to the objective. This primarily manifests in lower signal-to-noise ratio (SNR) of gradient signals, misalignment of update directions with true credit allocation, and large updates being absorbed by clipping/constraining mechanisms without resulting in effective progress. When such ineffective updates accumulate during training, they lead to systemic consequences in optimization dynamics, capability generalization, memory stability, and engineering costs. Below is an easy-to-understand section-wise discussion.

E.1. Optimization Dynamics: Convergence Degradation and Training Instability

(1) Effective Signal Drowned by Noise: Increased Gradient Variance and Decreased Sample Efficiency. Ineffective updates essentially inject task-unrelated noise into the parameters, reducing the proportion of effective gradients for the same computational budget. This directly leads to: (i) slower convergence; (ii) significantly more tokens and optimization steps required to reach the same performance threshold; (iii) more jittery training curves and prolonged plateaus.

(2) Adaptive Optimizer Statistics Polluted: Momentum and Second Moment Estimates Shifted. For adaptive optimizers like Adam/AdamW/Adafactor, parameter updates depend on historical first-order momentum and second-moment estimates. Long-term accumulation of ineffective gradients causes the momentum direction and scaling factors to be dominated by *incorrect gradient statistics*, leading to improper scaling of genuine gradients. This manifests as the need for smaller learning rates for stability, increased sensitivity to random seeds, and more frequent training oscillations.

(3) Clipping and Trust Region Mechanisms Overtriggered: Effective Updates Undermined. In PPO/GRPO/GSPO-style objectives, importance ratio anomalies and policy drift trigger clipping or KL/trust region constraints. When ineffective updates lead to more extreme importance ratio distributions, clipping mechanisms activate more frequently, causing *effective signals to be undermined as well*, forming a negative feedback loop: "the more updates, the more clipping; the more clipping, the less signal."

E.2. Capabilities and Generalization: Mode Collapse, Reward Hacking, and Generalization Degradation

(1) Entropy/Mode Collapse: Exploration and Diversity Decline. When training updates repeatedly reinforce surface patterns related to rewards but unrelated to true causal steps (e.g., fixed templates, surface formats, redundant phrasing), the model gradually shrinks to a few high-probability patterns. This leads to a decline in output diversity, and significantly impairs exploration and branching search abilities for long-chain reasoning.

(2) Reward Hacking Becomes Easier. Ineffective updates push the model towards "easier paths to high rewards" that do not genuinely solve the task, such as catering to scorer preferences, excessive explanation, or formatted outputs. This often leads to improvements in offline metrics but a decline in real task quality, particularly evident in out-of-distribution (OOD) evaluations.

(3) Pseudo-Feature Overfitting: Transferable Reasoning Operator Learning Blocked. When credit assignment is incorrect, the model is more likely to learn dataset biases, prompt triggers, or surface correlations rather than transferable reasoning operators. This results in apparent stability on training or same-distribution validation, but significantly worsens cross-task transfer and robustness.

E.3. Memory Stability: Catastrophic Forgetting and Capability Drift

(1) Catastrophic Forgetting: Existing Capabilities Damaged by Ineffective Perturbations. Ineffective updates apply continuous perturbations to many parameters, disrupting the structural integrity of existing capabilities, particularly those related to "fragile balances" such as language fluency, factual consistency, and alignment behavior. Empirically, this often manifests as an increase in some benchmarks while seemingly unrelated capabilities degrade without explanation.

(2) Capability Drift and Non-reproducibility: High Sensitivity to Random Seeds/Batches. When training is dominated by ineffective updates, the optimization trajectory behaves more like a random walk with momentum, leading to significantly different results for the same setup under different random seeds, poor version stability, frequent regression test failures, and increased engineering maintenance costs.

E.4. Resources and Engineering: Exploding Costs and Increased Difficulty of Subsequent Corrections

(1) Diminishing Marginal Returns: Less Effective Progress for the Same Computational Budget. The accumulation of learning tax is equivalent to consuming expensive online sampling tokens and optimization steps on updates with no net gain, significantly lowering sample efficiency and increasing training costs.

(2) Subsequent Alignment/Safety Corrections More Difficult: Stronger Pullback Efforts Needed. When a model is pushed away from its original parameter basin by ineffective updates, subsequent corrections using SFT/preference alignment data require higher training intensities, leading to new side effects (e.g., overfitting to alignment data, further forgetting of basic capabilities).

E.5. Behavioral Level: Long-range Reasoning Inconsistency and Decreased Self-Correction Ability

(1) Loss of Reasoning Chain Consistency: Intermediate Steps More Likely to Contradict Each Other. Ineffective updates disrupt token-level/step-level consistency constraints, causing the model to exhibit more logical leaps, contradictions, or unnecessary reasoning branches in long-chain reasoning.

(2) Difficulty Forming Self-Correction Loops: Same Errors Repeatedly Occur. If updates cannot be precisely attributed to erroneous tokens/steps, the model struggles to establish a stable "error detection–correction" mechanism, causing the same errors to repeatedly occur after training. This leads to processes being untrustworthy and difficult to eliminate with minimal additional training.

Summary. In conclusion, learning tax is not merely a waste of computational resources; its long-term accumulation contaminates optimizer statistics, triggers clipping feedback, induces entropy collapse and reward hacking, exacerbates catastrophic forgetting, and ultimately leads to a decline in model capability, generalization, and version stability. This phenomenon provides a unified mechanism perspective for understanding training instability and low sample efficiency

in long-range reasoning scenarios and offers a motivation for the design of process-level credit allocation or intra-group consistent weighting strategies.

F. Why the Symbol Asymmetry of Clipping Disrupts Intra-Group Cancellation

Corollary F.1 (GRPO-style clipping can cause intra-group cancellation failure outside the clipping range)**.** *Consider the commonly used clipped surrogate in GRPO (or its equivalent reformulation), where the piecewise selection is determined by the sign of the advantage: Let* $\bar{w} = \mathrm{clip}(w, 1-\varepsilon, 1+\varepsilon)$*, then for scalar weight* w *and advantage* A*, the "sign-sensitive" equivalent form is:*

$$\min(wA, \bar{w}A) = \begin{cases} A \cdot \min(w, \bar{w}), & \text{if } A \geq 0, \\ A \cdot \max(w, \bar{w}), & \text{if } A < 0. \end{cases}$$

In the case of shared context-token events, if there are two trajectories i, j *in the group such that* $w_{i,t^\star} = w_{j,t^\star} = w$ *and* $\widehat{A}_i > 0, \widehat{A}_j < 0$*, and* $w \notin [1-\varepsilon, 1+\varepsilon]$ *causing the two trajectories to enter* different branches *of the piecewise operator (for example, one trajectory takes* w*, the other takes* $\bar{w}$*), then the effective coefficients for the shared tokens within the group will no longer be consistent, which can result in the failure of intra-group cancellation and cause non-zero drift.*

For example, with $G = 2$ and $\widehat{A}_1 = -A, \widehat{A}_2 = +A$:

- If $w > 1+\varepsilon$, then $\bar{w} = 1+\varepsilon$. For positive advantage samples, we have $\min(wA, \bar{w}A) = \bar{w}A$, and for negative advantage samples, $\min(w(-A), \bar{w}(-A)) = w(-A)$, so the intra-group aggregation coefficient for the shared token in this direction is

$$\widehat{A}_1 \cdot w + \widehat{A}_2 \cdot \bar{w} = (-A)w + (A)(1+\varepsilon) = A\big((1+\varepsilon) - w\big) \neq 0 \quad (w \neq 1+\varepsilon). \tag{75}$$

- If $w < 1-\varepsilon$, then $\bar{w} = 1-\varepsilon$. For positive advantage samples, we have $\min(wA, \bar{w}A) = wA$, and for negative advantage samples, $\min(w(-A), \bar{w}(-A)) = \bar{w}(-A)$, so the intra-group aggregation coefficient is

$$\widehat{A}_1 \cdot \bar{w} + \widehat{A}_2 \cdot w = (-A)(1-\varepsilon) + (A)w = A\big(w - (1-\varepsilon)\big) \neq 0 \quad (w \neq 1-\varepsilon). \tag{76}$$

In conclusion, *clipping only maintains intra-group cancellation for shared tokens when* $w \in [1-\varepsilon, 1+\varepsilon]$*; once* w *exceeds the clipping range, the symbol asymmetry of clipping causes the effective coefficients for the shared token within the group to no longer be interchangeable, thus producing non-zero drift and learning tax.*

G. Non-zero Drift of Learning Tax in Statistical Significance: From Strict to Expectation-Based Cancellation Lower Bound

Corollary G.1 (Non-commutative weights induce non-zero expected drift on general tokens)**.** *Consider a class of tokens* $\mathcal{C}$ *that are* weakly related *to terminal rewards and appear frequently (e.g., template/functional words), and consider their "statistical cancellation" target during training:*

$$\mathbb{E}\Big[\widehat{A} \mid y \in \mathcal{C}\Big] \approx 0.$$

If the effective weight of tokens within a group, $\omega_{i,t}(\tau_i; \theta)$*, destroys commutativity, such that under the condition* $y_{i,t} \in \mathcal{C}$*,* $\omega_{i,t}$ *is systematically correlated with the group comparison signal* $\widehat{A}_i$*, i.e.,*

$$\mathrm{Cov}\Big(\widehat{A}_i, \nabla_\theta \omega_{i,t}(\tau_i; \theta) \;\Big|\; y_{i,t} \in \mathcal{C}\Big) \neq 0,$$

then the aggregated group gradient for this class of tokens will be strictly non-zero in expectation:

$$\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i \, \nabla_\theta \omega_{i,t}(\tau_i; \theta) \;\middle|\; y_{i,t} \in \mathcal{C}\right] \neq \mathbf{0},$$

thus producing an accumulative reward-irrelevant drift (learning tax) during training, consistent with the KL drift conclusion in Proposition 3.1.

Explanation (in one sentence). This corollary characterizes a typical scenario when "context is difficult to fully match" in practice: strict cancellation rarely occurs, but if weights are commutative, statistical averaging can approximate cancellation, making the learning tax smaller; however, once non-commutative and correlated with $\widehat{A}$, general tokens will exhibit sustained drift.

H. Symmetric Clipping: Fixing the Sign Asymmetry in GRPO Clipping (Restoring Group-wise Commutativity)

This appendix provides a **symmetric clipping** formulation to fix the **sign asymmetry** in GRPO-type clipped surrogates discussed in Appendix F (other GRPO variants like DAPO/DCPO/SSPO can be similarly adjusted): The standard GRPO objective triggers $\min$ / $\max$ branches for $A > 0$ and $A < 0$, respectively, which causes different *effective coefficients* even for identical token ratios r within a group, thus breaking the intra-group cancellation for shared tokens.

Recap: Asymmetry arises from "$\min$ and A sign coupling". For scalar ratio r, advantage A, and clipped ratio $\bar{r} = \mathrm{clip}(r, 1-\varepsilon, 1+\varepsilon)$, the standard surrogate is written as

$$\mathcal{L}_{\mathrm{ppo}}(r, A) \triangleq \min(rA,\ \bar{r}A). \tag{77}$$

Its equivalent piecewise form (see Appendix F) is:

$$\min(rA, \bar{r}A) = \begin{cases} A \cdot \min(r, \bar{r}), & A \geq 0, \\ A \cdot \max(r, \bar{r}), & A < 0, \end{cases} \tag{78}$$

This piecewise mapping explicitly depends on the sign of A, so when both positive and negative advantages exist within a group, the same r might enter different branches for different trajectories, thus breaking exchangeability and cancellation.

H.1. Symmetric Clipping: Decoupling Effective Weights from Advantage Sign

To restore commutativity for shared tokens, we rewrite the clipped surrogate in a **sign-independent** form:

$$\mathcal{L}_{\mathrm{sym}}(r, A) \triangleq A \cdot \phi(r), \qquad \phi(r) \triangleq \mathrm{clip}(r, 0, 1+\varepsilon). \tag{79}$$

That is, **for all samples, regardless of the sign of A, the same clipping operator $\phi(r)$ is applied.** In the token-factorized GRPO/PPOLike objectives, replace the original token-wise $r_{i,t}$ with $\phi(r_{i,t})$.

Symmetric Clipping Version of GRPO-fix (token-factorized). Let the group-wise advantage satisfy the zero-mean constraint $\sum_{i=1}^{G} \widehat{A}_i = 0$, and define the token-level ratio as $r_{i,t}(\theta) = \frac{\pi_\theta(a_t^{(i)} | h_t^{(i)})}{\pi_{\theta_{\mathrm{old}}}(a_t^{(i)} | h_t^{(i)})}$, then the symmetric clipping GRPO objective is

$$\mathcal{J}_{\text{GRPO-SymClip}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{T_i} \phi\big(r_{i,t}(\theta)\big)\ \widehat{A}_i\right], \qquad \phi(r) = \mathrm{clip}(r, 0, 1+\varepsilon). \tag{80}$$

The gradient (linear segment, ignoring subgradients at the breakpoints) is

$$\nabla_\theta \mathcal{J}_{\text{GRPO-SymClip}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{T_i} \widehat{A}_i\ \phi\big(r_{i,t}(\theta)\big)\ \nabla_\theta \log \pi_\theta\big(a_t^{(i)} \mid h_t^{(i)}\big)\right], \tag{81}$$

where $\phi(\cdot)$ is treated as a pointwise clipping scalar weight (in common implementations, subgradients can be used at breakpoints, or zero-measure sets can be ignored).

H.2. How It Fixes "Shared Token Cancellation"

Corollary H.1 (Symmetric Clipping Restores Intra-group Cancellation for Shared Tokens (Minimal Structural Check))**.** *Fix input x and time step $t^\star$, and consider event $\mathcal{E}_{t^\star}$: At this step, trajectories in the group share the same context-token pair $(h^\star, a^\star)$, so $r_{i,t^\star}(\theta) = r^\star(\theta)$ holds for all i, and the corresponding score function $\nabla_\theta \log \pi_\theta(a^\star \mid h^\star)$ is consistent*

across the group. If the group-wise advantage satisfies $\sum_{i=1}^{G} \widehat{A}_i = 0$*, then under the symmetric clipping objective* (80)*, the aggregated gradient for this shared token strictly cancels:*

$$\sum_{i=1}^{G} \widehat{A}_i \, \phi\big(r_{i,t^\star}(\theta)\big) \, \nabla_\theta \log \pi_\theta(a^\star \mid h^\star) = \phi\big(r^\star(\theta)\big) \Big(\sum_{i=1}^{G} \widehat{A}_i \Big) \nabla_\theta \log \pi_\theta(a^\star \mid h^\star) = \mathbf{0}. \tag{82}$$

Explanation. The key to this conclusion is that $\phi(\cdot)$ no longer depends on the sign of $\widehat{A}_i$, so the *effective weight* for shared tokens maintains commutativity within the group, thus enabling strict cancellation triggered by zero-mean advantages. This precisely fixes the source of exchangeability-breaking in the standard clipped surrogate (78).

H.3. Relation to Standard GRPO Surrogate and Cost

Differences. Symmetric clipping (79) adopts a more conservative update approach, resulting in slower updates but more stability (eliminating the associated learning tax).

Engineering Implementation (Recommended Minimal Change). If your current implementation is in the token-factorized GRPO/PPOLike form, simply replace the per-token effective weight from

$$\min(r_{i,t}\widehat{A}_i, \ \mathrm{clip}(r_{i,t})\widehat{A}_i)$$

with

$$\widehat{A}_i \cdot \mathrm{clip}(r_{i,t}, 0, 1+\varepsilon),$$

to implement the symmetric clipping objective (80).

I. Implementation Details

Models and Context Length. We configure **Qwen3-32B** with a context length of **32k** tokens, and **Qwen3-Next-80B-A3B-Thinking** with a context length of **256k** tokens. Inference is performed using the **vLLM** engine (version **0.11.2**).

Hardware. Experiments are conducted on **32 NVIDIA A800 (80GB)** GPUs.

Optimization Hyperparameters. The training hyperparameters are set as follows:

- Initial learning rate: 8×10^{-7};
- Learning rate schedule: cosine decay, with a minimum learning rate ratio of 0.2;
- Warmup: linear warmup covering 3% of the total training steps;
- Entropy regularization coefficient: $\beta = 0$;
- Rollouts: 32 trajectories (rollouts) are sampled for each input;
- Mini-batch size: 32.

J. Supplementary Experimental Setup

Inference Settings. The decoding parameters for Qwen3-32B are Temperature=0.6, TopP=0.95, TopK=20, MinP=0; for Qwen3-Next-80B-A3B-Instruct, the decoding parameters are Temperature=0.7, TopP=0.8, TopK=20, MinP=0. All results are compared under the same decoding settings.

K. Bias of Min-Replace and the Breakdown of Importance Sampling Unbiasedness: Effects, Bounds, and Testable Predictions

This appendix fully addresses a key question, both in terms of implementation and theory:

> **Does applying Min-Replace (taking the minimum and broadcasting) within a group break the unbiasedness of importance sampling (IS)? What are the effects? Does it lead to "wrong direction/backward updates"? When can the bias be ignored?**

The conclusions are summarized as follows (valid under the default implementation assumptions of this paper: **stop-gradient for group-wise transformation coefficients**):

1. **IS unbiasedness is indeed broken:** The estimator produced by Min-Replace no longer corresponds to the unbiased gradient of the original sequence-coupled IS objective; it optimizes a more *conservative* surrogate.
2. **No "backward updates" (sign flip):** Under stop-gradient, Min-Replace only performs *proportional shrinkage* on each trajectory's modulation coefficient, without reducing probabilities that should increase (the sign of updates determined by $\widehat{A}_i$ remains unchanged).
3. **The main effect is bias-variance tradeoff:** Min-Replace strongly suppresses the dominance of large ratio/tail samples within the group, significantly reducing variance, decreasing the over-triggering of clipping/KL constraints, and improving stability. The cost is introducing bias and smaller effective step sizes, potentially slowing convergence and, in extreme cases, shifting the optimal point (more strongly "pulling back" toward the old policy).
4. **When can the bias be ignored:** When training is constrained within a small trust region such that the group-wise $s_i(\theta)$ are close to each other (e.g., $|\log s_i| \le \delta$ and δ is small), the bias upper bound vanishes as $\delta \to 0$, and Min-Replace's main effect can be approximated as "robust variance reduction/implicit trust region."

K.1. Setup: Sequence-Coupled Weights and Min-Replace (using the DFPO-based GSPO as an example)

Recall Appendix B: We begin by rewriting the clipped surrogate of GSPO strictly equivalently in the form "first symbol-sensitive clipping of weights, then multiply by advantage":

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i \, \bar{s}_i(\theta)\right], \tag{83}$$

where $\bar{s}_i(\theta)$ is the post-clip weight defined in (47), and $\bar{s}_i(\theta) > 0$ holds by definition.

The Min-Replace transformation in DFPO acts on $\bar{\mathbf{s}}$ as follows:

$$\bar{s}_{\min}(\theta) \triangleq \min_{j\in\{1,\dots,G\}} \bar{s}_j(\theta), \qquad \widetilde{s}_i(\theta) \triangleq \bar{s}_{\min}(\theta) \;\; \forall i, \tag{84}$$

and stop-gradient is applied to $\widetilde{s}_i(\theta)$ during backpropagation (i.e., treating $\widetilde{s}_i$ as a constant coefficient within the group, not differentiating through the transformation operator; see the implementation conventions in Appendix D).

K.2. Gradient Form: Min-Replace Does Not Reverse Updates (Sign Preservation), Only Performs Conservative Shrinkage

Consider the linear segment (ignoring subgradient details at the breakpoints; this does not affect sign conclusions). Under stop-gradient, the DFPO gradient can be written as (structurally identical to (68)):

$$\nabla_\theta \widetilde{\mathcal{J}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \widetilde{s}_i \, \widehat{A}_i \cdot \sum_{t=1}^{|y_i|} \alpha_{i,t} \, \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right], \tag{85}$$

where $\alpha_{i,t} \geq 0$ are non-negative coefficients (e.g., $\alpha_{i,t} = 1/|y_i|$).

The gradient of the baseline (without Min-Replace) post-clip objective (83) under the stop-gradient assumption is written as:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \bar{s}_i(\theta)\, \widehat{A}_i \cdot \sum_{t=1}^{|y_i|} \alpha_{i,t}\, \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right]. \tag{86}$$

Since $\widetilde{s}_i(\theta) = \bar{s}_{\min}(\theta) \leq \bar{s}_i(\theta)$ and both are positive, we define the shrinkage ratio for each trajectory:

$$\phi_i(\theta) \triangleq \frac{\widetilde{s}_i(\theta)}{\bar{s}_i(\theta)} = \frac{\bar{s}_{\min}(\theta)}{\bar{s}_i(\theta)} \in (0, 1]. \tag{87}$$

Thus, for any trajectory i, its modulation coefficient satisfies:

$$\widehat{A}_i\, \widetilde{s}_i(\theta) = \phi_i(\theta) \cdot \widehat{A}_i\, \bar{s}_i(\theta), \qquad \phi_i(\theta) \in (0, 1]. \tag{88}$$

Theorem K.1 (No Reverse Updates: Min-Replace Only Performs Proportional Shrinkage)**.** *Under the stop-gradient assumption, for any i and any token t, the coefficient of the score function term $\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})$ for the trajectory after Min-Replace preserves the sign with the baseline:*

$$\operatorname{sign}\Big(\widehat{A}_i\, \widetilde{s}_i(\theta)\Big) = \operatorname{sign}\Big(\widehat{A}_i\, \bar{s}_i(\theta)\Big) = \operatorname{sign}\Big(\widehat{A}_i\Big), \tag{89}$$

Hence, Min-Replace does not cause "reverse updates" (no sign flip for updates determined by the advantage of individual trajectories), and its effect is equivalent to applying more conservative effective step sizes for trajectories with non-minimum weights.

Proof (Simplified). Since $\bar{s}_i(\theta) > 0$, $\widetilde{s}_i(\theta) = \bar{s}_{\min}(\theta) > 0$, and from (88), we see that $\widehat{A}_i\, \widetilde{s}_i(\theta)$ is a positive scalar multiple of $\widehat{A}_i\, \bar{s}_i(\theta)$, so the sign is preserved. □

K.3. Where Does the Bias Come From: Min-Replace Breaks IS Unbiasedness and Alters the Optimization Objective

Although there is no reverse update, Min-Replace **does introduce bias**: it no longer corresponds to the Radon–Nikodym derivative form of the original IS weights, so it generally does not satisfy "unbiased transport to a target distribution."

It is more intuitive to view this from the perspective of the "stop-gradient gradient estimator." Define the trajectory-level vector:

$$\mathbf{G}_i(\theta) \triangleq \sum_{t=1}^{|y_i|} \alpha_{i,t}\, \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}), \tag{90}$$

Then, the single-step gradient estimates for the baseline and Min-Replace can be written as:

$$\widehat{g}_{\text{base}} = \frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i\, \bar{s}_i(\theta)\, \mathbf{G}_i(\theta), \qquad \widehat{g}_{\min} = \frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i\, \bar{s}_{\min}(\theta)\, \mathbf{G}_i(\theta). \tag{91}$$

The expected difference between them (the bias vector) is:

$$\operatorname{Bias}(\theta) \triangleq \mathbb{E}[\widehat{g}_{\min}] - \mathbb{E}[\widehat{g}_{\text{base}}] = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i\big(\bar{s}_{\min}(\theta) - \bar{s}_i(\theta)\big)\mathbf{G}_i(\theta)\right]. \tag{92}$$

Since $\bar{s}_{\min} - \bar{s}_i \leq 0$, this bias is generally non-zero, implying that **Min-Replace optimizes a "more conservative, closer to the old policy" surrogate, not the original IS objective.**

Additionally, an upper bound without extra assumptions can be given, characterizing how the bias grows as the "group-wise weight dispersion" increases:

$$\big\|\operatorname{Bias}(\theta)\big\| \leq \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} |\widehat{A}_i|\, \big(\bar{s}_i(\theta) - \bar{s}_{\min}(\theta)\big)\, \big\|\mathbf{G}_i(\theta)\big\|\right]. \tag{93}$$

Thus, when the $\bar{s}_i$ within a group are more dispersed (especially with long-tail large ratios), the bias of Min-Replace becomes more pronounced; when they are close (within a trust region), the bias is smaller.

K.4. Bias-Variance Tradeoff: Why Min-Replace is Generally More Stable and Less Prone to Over-triggering Clipping Negative Feedback

The direct structural effect of Min-Replace is: it changes the group-wise modulation coefficients from $\{\widehat{A}_i \bar{s}_i\}$ to $\{\widehat{A}_i \bar{s}_{\min}\}$, eliminating the random modulation induced by the dispersion of $\bar{s}_i$ within the group. This leads to two typical benefits:

(1) Decreased asymmetric modulation within the group (aligned with the mechanism measure in the main text). The main text defines

$$\mathrm{Asym}(t) = \mathrm{Var}_{i\in\{1,\dots,G\}}\Big(w_i(\tau_i;\theta)\,\widehat{A}_i\Big) \tag{94}$$

to measure the strength of "difficult to cancel out" effects on shared/similar tokens. Under Min-Replace, if we view trajectory-level modulation factors as $w_i\widehat{A}_i$, then w_i is forced to be constant within the group ($\bar{s}_{\min}$), which significantly reduces the variance, making it easier to recover (or approximate) cancellation structures on shared/high-frequency token subspaces, thus reducing learning tax.

(2) Tail ratio dominance is suppressed, clipping/KL constraints trigger less frequently and with less "false positives." In PPO/GSPO-type objectives, large-ratio samples often cause the surrogate to enter the clipping region, frequently triggering constraints and forming the negative feedback of "more updates lead to more clipping, more clipping leads to no signal." Min-Replace forces all trajectory modulations to the minimum group weight, applying stronger shrinkage to large-ratio samples, which typically reduces the over-triggering of clipping/KL constraints and improves stability.

The cost is the bias shown in (92): the updates become more conservative, effective step sizes are smaller, potentially slowing convergence or shifting the optimal point closer to the old policy.

K.5. When Can the Bias Be Ignored: A Sufficient Condition for the Trust Region (Can Explain Ablation Phenomena)

When training occurs within a small trust region, so that the post-clip weights within the group are close to each other, the bias is controlled. For example, if there exists $\delta > 0$ such that for all trajectories within the same group:

$$\big|\log \bar{s}_i(\theta)\big| \le \delta, \tag{95}$$

then $\bar{s}_i(\theta) \in [e^{-\delta}, e^{\delta}]$, and we see that

$$0 \le \bar{s}_i(\theta) - \bar{s}_{\min}(\theta) \le e^{\delta} - e^{-\delta}, \qquad \phi_i(\theta) = \frac{\bar{s}_{\min}}{\bar{s}_i} \in [e^{-2\delta}, 1]. \tag{96}$$

Substituting this into the bias upper bound (93), we observe that: **when δ is small enough (ratios are strictly limited), the upper bound of Min-Replace's bias tends to 0 as $\delta \to 0$.** At this point, the main effect of Min-Replace can be approximated as "variance reduction + implicit smaller step size/stronger trust region," rather than "severely rewriting the objective."

K.6. Implications for Ablations and Experiments in This Paper: Which Phenomena Correspond to Bias and Which to Structural Fixes?

Combining the ablation results from Table 3:

- **Removing stop-gradient (DFPO no stop-grad) leads to performance degradation:** This is not due to the "IS bias" itself, but because allowing group transformations to participate in backpropagation introduces additional gradient coupling and instability, breaking the required implementation assumption of "treating transformations as group-wise control variables," thus introducing new exchangeability-breaking.
- **Replacing group-wise normalization with global scaling (DFPO scale by 0.5) still performs significantly worse:** This indicates that the gain is not solely due to "smaller effective step sizes/more conservative updates" (bias side), but crucially arises from Min-Replace's elimination of weight dispersion within the *group*, reducing Asym and restoring (or approximating) cancellation structures on shared/high-frequency token subspaces, thus reducing learning tax (structural fix side).

Summary. Min-Replace breaks IS unbiasedness and thus introduces bias; however, under the stop-gradient assumption, it does not cause ”reverse updates” but applies proportional shrinkage to non-minimal-weight trajectories. The core benefits manifest as reduced variance, decreased asymmetric modulation within the group, and reduced learning tax; the cost is a more conservative surrogate objective and potentially slower convergence. These conclusions align precisely with the mechanistic analysis in this paper regarding the necessary conditions for exchangeability-cancellation.