ONE STEP, ONE LEAD: MITIGATING HIGHER-ORDER INTERFERENCE IN MULTI-DOMAIN REINFORCEMENT LEARNING VIA CROSS-STEP CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Although Reinforcement Learning with Verifiable Rewards (RLVR) provides a principled post-training paradigm for cultivating complex reasoning abilities of Large Language Models (LLMs), jointly training one foundation model across heterogeneous domains usually introduces multi-domain trade-offs and destabilizes the overall policy optimization. While existing approaches typically diagnose cross-domain interference via gradient alignment or curvature proxies, they usually overlook interference that accumulates across consecutive updates—where seemingly non-conflicting gradients can still overwrite prior policy gains. We show that these cross-step dynamics are directly tractable between adjacent checkpoints through token log-probability footprints, bypassing costly curvature approximations. Building on this observation, we propose OSOL, which organizes each mixed-domain update around a single focus domain and uses the preceding checkpoint footprint to identify rebound-prone tokens and apply a rank-based, adaptively scaled correction. Our analysis shows that this correction contracts the targeted negative-to-positive backtracking component, while controlled studies show that cross-step backtracking is more strongly associated with subsequent task damage than within-step gradient diagnostics and that checkpoint footprints identify future rebound more effectively than Hessian-based proxies. Across Qwen3-8B-Base and Qwen3-30B-A3B, OSOL achieves the best four-domain macro average on both backbones, while ranking first or second in most evaluations. On Qwen3-30B-A3B, it reaches a domain-macro average of 0.4822, outperforming the strongest baseline by 5.7%. Ablations support prioritizing one domain per step, using checkpoint drift to locate rebound-prone tokens, and correcting them at the token level. Code is available at `https://anonymous.4open.science/r/anonymous-osol-83DF/`.

1 INTRODUCTION

Reinforcement learning (RL) has become a central component of LLM post-training, substantially improving reasoning and task-solving capabilities across individual domains (Shao et al., 2024; Yu et al., 2026). For general-purpose models, however, post-training must improve multiple capabilities jointly, making multi-domain RL essential for broad and balanced capability development. In practice, jointly training heterogeneous domains often creates trade-offs: improving one capability can come at the expense of others and may even destabilize optimization (Cai et al., 2026; Yang et al., 2026b; Li et al., 2025). As Figure 1 shows, mixed-domain GRPO can substantially underperform its single-domain counterparts on individual tasks. This raises a central question: how do updates from different domains interfere during training, and how can we control this interference to improve multi-domain RL?

Existing approaches mainly analyze cross-domain interference within a single optimization step, through first-order gradient conflicts or higher-order gradient interactions (Yu et al., 2020; Liu et al., 2021; Liang et al., 2026; Yang et al., 2026a). However, these within-step diagnostics can miss interference that unfolds across updates: a later update may overwrite gains from the previous one even when the two appear locally non-conflicting. As shown in Figure 2, gradient alignment measured before the second update is only weakly related to its subsequent task damage, while expanding

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

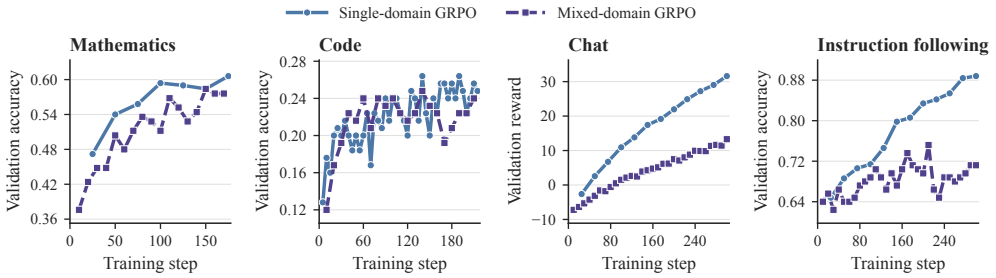

Figure 1: **Mixed-domain GRPO underperforms single-domain training.** Validation trajectories show clear gaps in Mathematics, Chat, and Instruction Following, with a smaller gap in Code.

the Hessian proxy to more parameters yields limited gains in predicting token rebound at substantially higher computational cost. These findings motivate a trajectory-level view of multi-domain interference that directly examines the realized interaction between consecutive updates.

We capture this interaction through token log-probability footprints: the changes an update induces in the log-probabilities of a fixed set of realized response tokens. When the next update moves these log-probabilities back in the opposite direction, it reveals cross-step output backtracking. We show that the inner product of adjacent token footprints locally recovers the mixed second-order interaction between consecutive updates under the empirical output Fisher. Building on this observation, we propose OSOL. At each step, OSOL selects one focus domain and uses only the preceding checkpoint footprint to identify tokens most likely to rebound. It then applies a rank-based, adaptively scaled correction to their GRPO coefficients. In this way, OSOL turns the token-level footprint left by the previous update into an online control signal, without auxiliary supervision or explicit higher-order differentiation. We make three contributions:

1. **A cross-step view of interference in multi-domain RL.** We show that within-step gradient diagnostics can miss interference realized across consecutive policy updates. Adjacent policy checkpoints induce token log-probability footprints that directly expose this interaction in policy output space and locally recover its mixed second-order structure under the empirical output Fisher.
2. **Checkpoint-guided cross-step control.** We propose OSOL, which prioritizes one domain per step, uses checkpoint drift to locate rebound-prone tokens, and applies a token-level correction within the standard GRPO objective. We further show that this correction contracts the targeted negative-to-positive backtracking component under a local response condition.
3. **Strong multi-domain gains and mechanism evidence.** OSOL ranks first in domain-macro average on both Qwen3-8B-Base and Qwen3-30B-A3B, whereas the strongest baseline differs between the two backbones; on Qwen3-30B-A3B, it improves over the strongest baseline by 5.7%. There, the gain reaches 17.4% in code generation while also improving mathematics by 1.1% over the strongest domain-specific baselines. A matched ablation further shows that adding checkpoint-history correction improves the domain-macro average by 3.8% over the same focus structure without history correction. Controlled studies further support the cross-step backtracking signal and checkpoint-guided token correction.

2 Related Work

Recent studies of multi-domain RL post-training report asymmetric transfer, domain-specific degradation, uneven gradient contributions across tasks, and sensitivity to data composition and training order (Cheng et al., 2026; Li et al., 2025; Wu et al., 2026; Yang et al., 2026b; Hu et al., 2026; Lu et al., 2026). Existing optimization methods mainly address these issues by balancing task contributions or coordinating gradients within a single optimization step. Adaptive weighting methods adjust task weights according to gradient magnitudes or convergence rates (Chen et al., 2018; Gong et al., 2024), while multi-objective approaches seek Pareto-efficient, bargaining-based, or balanced-progress solutions (Sener & Koltun, 2018; Navon et al., 2022; Liu et al., 2023). PCGrad and CAGrad instead modify conflicting task gradients before aggregation (Yu et al., 2020; Liu et al., 2021). Re-

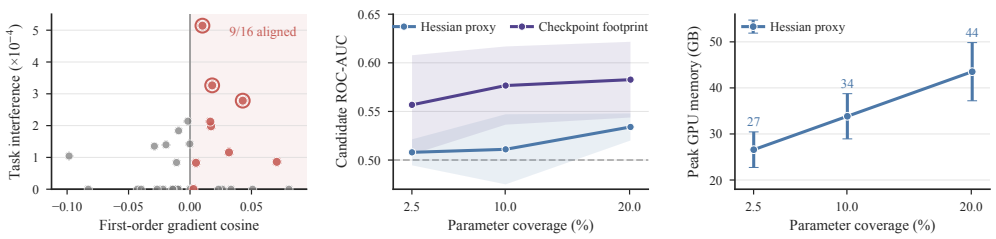

Figure 2: **Gradient alignment before the second update does not reliably indicate its subsequent task damage (a), while Hessian-based proxies are costly (b, c).**

cent LLM-specific methods extend these ideas through module-level gradient surgery, adaptive task weighting and sampling, and transfer-aware curricula (Cai et al., 2026; Ramesh et al., 2026; Yang et al., 2026c). Overall, these methods primarily regulate task contributions or gradient interactions within the current update, whereas our work focuses on interference that emerges across consecutive mixed-domain updates.

Three lines of work are particularly relevant to our cross-step perspective: higher-order optimization, local policy geometry, and checkpoint-based training analysis. Influence functions characterize the sensitivity of model predictions to training perturbations using inverse-Hessian information (Koh & Liang, 2017). More recent work studies cross-domain interactions through local curvature: local-perturbation analyses examine how a small update on one domain affects another through higher-order terms (Yang et al., 2026a), while CGPO induces curvature-aware cross-domain interactions by reevaluating each domain gradient after small perturbations from other domains, thereby approximating Hessian–gradient interactions without explicit Hessian computation (Liang et al., 2026). Natural-gradient and trust-region methods provide a complementary geometric perspective, using the Fisher metric and policy KL to measure local changes in the policy distribution (Amari, 1998; Kakade, 2001; Schulman et al., 2015; Martens, 2020). Related work on the empirical Fisher also emphasizes that it should not be interpreted as a generic replacement for an arbitrary objective Hessian (Kunstner et al., 2019). Finally, checkpoint-based influence methods show that stored checkpoints can reveal how training examples affect the optimization trajectory (Pruthi et al., 2020). In contrast to methods that infer cross-domain interaction from gradients or local parameter perturbations within the current optimization step, we directly characterize the interaction realized across consecutive policy updates through token-level output changes between adjacent checkpoints.

3 Cross-Step Dynamics

For a fixed set of realized response-token coordinates $\{(x_i, a_i)\}_{i=1}^n$, where x_i is the prefix context and a_i is the realized token, we work directly in the corresponding token log-probability coordinates. The footprint of a parameter update is the induced change in this log-probability vector. Consider

$$\boldsymbol{\theta}_{k-1} \xrightarrow{\boldsymbol{\delta}_{k-1}} \boldsymbol{\theta}_k \xrightarrow{\boldsymbol{\delta}_k^0} \boldsymbol{\theta}_{k+1}^0, \tag{1}$$

where $\boldsymbol{\delta}_{k-1} := \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}$ is the preceding realized update and $\boldsymbol{\delta}_k^0 := \boldsymbol{\theta}_{k+1}^0 - \boldsymbol{\theta}_k$ is the current base update before the OSOL correction. We analyze the two adjacent updates on the same token coordinates, making their interaction directly observable from checkpoint log-probability changes.

3.1 Output Geometry

Let $\mathcal{E}_k = \{(x_i, a_i)\}_{i=1}^N$ denote eligible response-token coordinates from the domain selected for cross-step control at iteration k, which we call the focus domain; Section 4 gives its role in the training update and Section 4.1 specifies the implementation-level eligibility rule. Let $\boldsymbol{\ell}_k(\boldsymbol{\theta}) := [\log \pi_{\boldsymbol{\theta}}(a_i \mid x_i)]_{i=1}^N$ collect the corresponding token log-probabilities. Define the output KL on these focus-domain token contexts

$$\mathcal{D}_k(\boldsymbol{\delta}) := \frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}_k}(\cdot \mid x_i) \,\|\, \pi_{\boldsymbol{\theta}_k + \boldsymbol{\delta}}(\cdot \mid x_i)) \,. \tag{2}$$

For $\boldsymbol{\varphi}_{i,a} := \nabla \log \pi_{\boldsymbol{\theta}_k}(a \mid x_i)$ and $\boldsymbol{\varphi}_i := \boldsymbol{\varphi}_{i,a_i}$, define

$$\overline{\boldsymbol{F}}_k := \frac{1}{N} \sum_i \mathbb{E}_{a \sim \pi_{\boldsymbol{\theta}_k}(\cdot|x_i)}[\boldsymbol{\varphi}_{i,a} \boldsymbol{\varphi}_{i,a}^\top], \qquad \boldsymbol{F}_k := \frac{1}{N} \sum_i \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^\top. \tag{3}$$

The first, which we call the output Fisher, averages score outer products over the full next-token distribution at each context. The second is the realized empirical output Fisher formed from the response tokens actually observed in the rollout (Amari, 1998; Martens, 2020; Kunstner et al., 2019). Throughout, $\overline{\boldsymbol{F}}_k$ describes the local KL geometry, whereas $\boldsymbol{F}_k$ is the realized metric appearing in the checkpoint identity below.

By the standard Fisher–KL identity, under the local regularity conditions in Appendix C.1,

$$\mathcal{D}_k(\boldsymbol{\delta}) = \tfrac{1}{2} \boldsymbol{\delta}^\top \overline{\boldsymbol{F}}_k \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3), \tag{4}$$

where $\overline{\boldsymbol{F}}_k = \nabla^2 \mathcal{D}_k(\mathbf{0})$ is the exact local metric induced by the output KL (Amari, 1998; Martens, 2020; Kakade, 2001; Schulman et al., 2015). Derivation and remainder details are given in Appendices C.3 and C.4.

On the same realized coordinates, define the preceding footprint and the current base footprint

$$\boldsymbol{u} := \boldsymbol{\ell}_k(\boldsymbol{\theta}_k) - \boldsymbol{\ell}_k(\boldsymbol{\theta}_{k-1}), \qquad \boldsymbol{d}^0 := \boldsymbol{\ell}_k(\boldsymbol{\theta}_{k+1}^0) - \boldsymbol{\ell}_k(\boldsymbol{\theta}_k). \tag{5}$$

Theorem 1 (Checkpoint token–Fisher duality)**.** *If $\|\boldsymbol{\delta}_{k-1}\|, \|\boldsymbol{\delta}_k^0\| = O(\eta)$, then*

$$\frac{1}{N} \boldsymbol{u}^\top \boldsymbol{d}^0 = \boldsymbol{\delta}_{k-1}^\top \boldsymbol{F}_k \boldsymbol{\delta}_k^0 + O(\eta^3). \tag{6}$$

Coordinatewise, $u_i d_i^0 < 0$ means that the current update reverses the preceding log-probability movement on token i, whereas $u_i d_i^0 > 0$ means that it continues that movement. The inner product of the preceding and current base footprints therefore exposes the mixed output interaction between the preceding realized update and the current base update under the empirical output Fisher. Both footprints are defined on the same realized token coordinates. The current base footprint $\boldsymbol{d}^0$ is used only for analysis, while the online controller uses only the preceding footprint $\boldsymbol{u}$. The coordinatewise expansion and explicit finite-step error are given in Appendices D.1–D.3.

Token Backtracking. The interaction has the exact path interpretation

$$\Gamma(\boldsymbol{u}, \boldsymbol{d}) := -\frac{1}{N} \boldsymbol{u}^\top \boldsymbol{d} = \frac{\|\boldsymbol{u}\|^2 + \|\boldsymbol{d}\|^2 - \|\boldsymbol{u} + \boldsymbol{d}\|^2}{2N}. \tag{7}$$

Hence $\Gamma > 0$ measures stepwise output motion that is canceled in the net displacement. Together with Theorem 1,

$$\Gamma(\boldsymbol{u}, \boldsymbol{d}^0) = -\boldsymbol{\delta}_{k-1}^\top \boldsymbol{F}_k \boldsymbol{\delta}_k^0 + O(\eta^3), \tag{8}$$

which gives the mixed output interaction a direct trajectory interpretation. Appendix D.5 decomposes the global quantity as $\Gamma = \mathrm{Rev} - \mathrm{Align}$, where $\mathrm{Rev} = \mathcal{B}^- + \mathcal{B}^+$ aggregates opposing token movements and Align aggregates same-direction movements. Under the sign-transfer condition introduced below, the nonpositive history correction used by OSOL has a monotone reduction guarantee for the negative-to-positive component; see Appendix F.1. We therefore define the controlled backtracking mass as

$$\mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}) := \frac{1}{N} \sum_{i: u_i < 0} (-u_i)[d_i]_+, \tag{9}$$

which captures negative-to-positive rebound: a token log-probability decreased by the preceding update and immediately increased by the next one.

3.2 Backtracking Control

OSOL applies a nonpositive residual on $\mathcal{C}_k := \{i \in \mathcal{E}_k : u_i < 0\}$. Let $\xi_i \le 0$ denote its effective post-clipping change to the policy coefficient; monotonicity of scalar clipping preserves this sign (Appendix E.1).

Lemma 1 (History-conditioned PPO penalty). *For $z_i := \log(\pi_{\boldsymbol{\theta}}(a_i \mid x_i)/\pi_{\boldsymbol{\theta}_k}(a_i \mid x_i))$, the token-level loss difference in a local neighborhood of the PPO anchor satisfies*

$$\mathcal{J}_i^{\mathrm{OSOL}}(z_i) - \mathcal{J}_i^0(z_i) = -e^{z_i}\xi_i, \qquad \partial_{z_i}(\mathcal{J}_i^{\mathrm{OSOL}} - \mathcal{J}_i^0)\big|_{z_i=0} = -\xi_i \geq 0. \tag{10}$$

Thus, locally, gradient descent adds a history-conditioned penalty against increasing the selected likelihood ratios. Appendix E.2 derives the result.

Let $\boldsymbol{\Phi}_{\mathcal{E}} := [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_N]$ and $\boldsymbol{K}_{\mathcal{E}} := \boldsymbol{\Phi}_{\mathcal{E}}^\top \boldsymbol{\Phi}_{\mathcal{E}}$, where $\boldsymbol{K}_{\mathcal{E}}$ is the token-gradient similarity matrix and captures cross-token coupling induced by shared model parameters. In the local policy-gradient model, the correction induces

$$\boldsymbol{h} := \boldsymbol{\Phi}_{\mathcal{E}}^\top \boldsymbol{\delta}_k^r = \frac{\eta}{|\mathcal{R}|}\boldsymbol{K}_{\mathcal{E}}\boldsymbol{\xi}, \tag{11}$$

where $\boldsymbol{h}$ is the correction footprint. Because all tokens share model parameters, a negative coefficient correction does not by itself guarantee a negative log-probability change for every selected token: cross-token interactions in $\boldsymbol{K}_{\mathcal{E}}$ can alter the local response. The following sufficient condition requires each selected token's own local response to dominate the aggregate cross-token effect.

Lemma 2 (Residual-to-output sign transfer). *Under the local dominance condition in Appendix E.4, $h_i \leq 0$ for the selected coordinates.*

Under this condition, the nonpositive coefficient correction transfers to a nonpositive output-space correction on the selected tokens. Appendix E.4 gives the exact condition and its extension to the locally preconditioned optimizer response.

Proposition 1 (Backtracking contraction). *If $h_i \leq 0$ on $\mathcal{C}_k$, then*

$$\Delta_k^{\mathrm{ctrl}} := \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0) - \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0 + \boldsymbol{h}) \geq 0. \tag{12}$$

For the corrected footprint $\boldsymbol{d}^{\mathrm{OSOL}}$ with $\|\boldsymbol{\delta}_k^r\| = O(\eta)$,

$$\mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^{\mathrm{OSOL}}) \leq \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0) - \Delta_k^{\mathrm{ctrl}} + O(\eta^3). \tag{13}$$

The proof and explicit finite-step remainder are in Appendices E.5 and E.6. Together, these results connect checkpoint-observable cross-step interaction to local contraction of the targeted backtracking component.

4 Method Design

OSOL turns the cross-step signal identified above into an online correction. Each update remains mixed-domain but designates one focus domain: all domains contribute to the base update, while only focus-domain response tokens receive history correction. The preceding checkpoint footprint on these tokens is then used to construct a nonpositive residual added to the standard GRPO coefficient.

4.1 Focus Weighting and Update History

For each prompt, standard GRPO samples a group of G responses. Let R_j denote the reward of response j. Its group-normalized trajectory advantage is

$$A_j^{\mathrm{g}} = \frac{R_j - \bar{R}}{\sqrt{\sigma_R^2 + \epsilon_{\mathrm{adv}}}}, \qquad \bar{R} = \frac{1}{G}\sum_{m=1}^{G} R_m, \tag{14}$$

where σ_R^2 is the within-group reward variance and $\epsilon_{\mathrm{adv}} > 0$ is a numerical stabilizer. The same trajectory advantage is assigned to every token in response j; below, A_i^{g} denotes this value for token i. At each step, the focus domain is sampled uniformly from the domains other than the preceding focus. We apply focus-dependent task weighting to the GRPO signal. Let $\mathrm{src}(i)$ denote the domain of token i. With focus and non-focus weights $w_{\mathrm{f}} > w_{\mathrm{nf}}$, the task-weighted base coefficient is

$$A_i^{\mathrm{base}} = c_i A_i^{\mathrm{g}}, \qquad c_i = \frac{w_{\mathrm{src}(i)}}{Z_k}, \qquad Z_k = \frac{w_{\mathrm{f}} N_k^{\mathrm{f}} + w_{\mathrm{nf}} N_k^{\mathrm{nf}}}{N_k^{\mathrm{f}} + N_k^{\mathrm{nf}}}, \tag{15}$$

where N_k^{f} and N_k^{nf} are focus- and non-focus-token counts. By construction, the token-weighted mean of c_i is one, so changing the focus redistributes weight without changing the average coefficient scale. We add the history residual after this weighting and do not multiply it by c_i again.

Let $\mathcal{E}_k$ contain current focus-domain response tokens with nonzero trajectory-level GRPO advantage. On each realized coordinate (x_i, a_i), compute

$$u_i = \log \pi_{\boldsymbol{\theta}_k}(a_i \mid x_i) - \log \pi_{\boldsymbol{\theta}_{k-1}}(a_i \mid x_i), \qquad \mathcal{C}_k = \{i \in \mathcal{E}_k : u_i < 0\}. \tag{16}$$

Thus $\boldsymbol{u}$ is the preceding footprint on the current rollout coordinates, not a prediction of the next update. Both checkpoints score the same (x_i, a_i) pairs, so the comparison requires no token matching or auxiliary model. A token enters $\mathcal{C}_k$ when its log-probability at the current prefix is lower under $\boldsymbol{\theta}_k$ than under $\boldsymbol{\theta}_{k-1}$. Under the sign-transfer condition in Lemma 2, this defines the negative-drift candidate set on which the correction monotonically suppresses negative-to-positive rebound (Appendix F.1).

4.2 Residual Construction

For $i \in \mathcal{C}_k$, let $\mathrm{rank}_0(u_i)$ be the ascending zero-based rank and set

$$p_i = \frac{\mathrm{rank}_0(u_i) + \frac{1}{2}}{|\mathcal{C}_k|}, \qquad \bar{r}_i = r_{\min} + \frac{\mathrm{clip}(p_i, p_{\mathrm{lo}}, p_{\mathrm{hi}}) - p_{\mathrm{lo}}}{p_{\mathrm{hi}} - p_{\mathrm{lo}}}(r_{\max} - r_{\min}), \tag{17}$$

with $\bar{r}_i = 0$ off $\mathcal{C}_k$. We use $(p_{\mathrm{lo}}, p_{\mathrm{hi}}, r_{\min}, r_{\max}) = (0.1, 1, -1, 0)$. With these values, the most negative decile receives the maximal penalty, while the remaining candidates are ranked progressively toward zero. For active, unsaturated rebound coordinates, the corresponding output-space ordering is locally optimal for allocating a fixed collection of penalty levels (Appendix F.2). Using ranks rather than raw drift magnitudes also reduces sensitivity to batch-to-batch scale changes and isolated extreme log-probability shifts.

Define the supported profile and its adaptive scale by

$$q_i := \bar{r}_i \mathbf{1}\{i \in \mathcal{C}_k\}, \tag{18}$$

$$\kappa_k := \tau \frac{\mathrm{std}_{\mathcal{E}_k}(A^{\mathrm{base}})}{\mathrm{std}_{\mathcal{E}_k}(q)}, \qquad r_i^{\mathrm{OSOL}} := \kappa_k q_i. \tag{19}$$

Whenever the statistics are nondegenerate,

$$\mathrm{std}_{\mathcal{E}_k}(r^{\mathrm{OSOL}}) = \tau\, \mathrm{std}_{\mathcal{E}_k}(A^{\mathrm{base}}). \tag{20}$$

The standard setting uses $\tau = 0.03$. Hence τ controls the residual relative to the current base-coefficient scale rather than to the raw scale of the preceding drift. This separates where to intervene, determined by drift rank, from how strongly to intervene, calibrated to the current base-coefficient scale. Numerical fallback and the code-level gate and mask are specified in Appendix F.3–F.4.

The final policy coefficient is

$$\widetilde{A}_i = \mathrm{clip}\big(A_i^{\mathrm{base}} + r_i^{\mathrm{OSOL}}, -3, 3\big). \tag{21}$$

With $\omega_i(\boldsymbol{\theta}) := \pi_{\boldsymbol{\theta}}(a_i \mid x_i)/\pi_{\boldsymbol{\theta}_{\mathrm{old}}}(a_i \mid x_i)$, it is optimized by the unchanged clipped surrogate (Schulman et al., 2017):

$$\mathcal{L}_{\mathrm{OSOL}}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{1}{|\mathcal{R}|}\sum_{i \in \mathcal{R}} \min\Big(\omega_i(\boldsymbol{\theta})\widetilde{A}_i, \mathrm{clip}(\omega_i(\boldsymbol{\theta}), 1-\epsilon_c, 1+\epsilon_c)\widetilde{A}_i\Big)\right]. \tag{22}$$

Thus, OSOL converts the preceding token drift into a tokenwise history-conditioned correction within the standard GRPO objective.

5 Experiments

5.1 Setup

Models and baselines. We evaluate OSOL on Qwen3-30B-A3B and Qwen3-8B-Base. On both backbones, we compare against the pre-RL base model, standard GRPO (Guo et al., 2025), GRPO+$\mathrm{KL}_{0.001}$ as a conservative-update baseline, MGS (Cai et al., 2026) as a first-order module-level gradient-surgery baseline, and CGPO (Liang et al., 2026) as a curvature-guided multi-domain optimization baseline. All methods follow the same evaluation protocol (details in Appendix F.5).

Table 1: **Main results on the Qwen3-30B-A3B backbone**, where OSOL achieves the best overall multi-domain performance. Best and second-best values are bolded and underlined.

Method	Math	Code	IF	Chat	Domain Avg.
Base model	0.5312	0.3310	**0.2100**	0.6392	0.4279
GRPO	0.5632	0.2603	0.2066	0.6900	0.4300
GRPO+$KL_{0.001}$	0.5538	0.2540	0.1900	**0.7271**	0.4312
MGS	<u>0.5757</u>	<u>0.3858</u>	0.2000	0.6625	<u>0.4560</u>
CGPO	0.5670	0.3835	<u>0.2067</u>	0.6541	0.4528
OSOL (ours)	**0.5822**	**0.4529**	0.1933	<u>0.7003</u>	**0.4822**

Training data. We construct a training dataset mixture covering four domains: mathematics, code generation, instruction following (IF), and open-ended chat (Chat). The mathematics and code examples are quality-filtered subsets of the NVIDIA Nemotron Post-Training Dataset v2 (Basant et al., 2025). For Chat, we use the WildChat-IF subset, which is derived from the Tulu 3 WildChat-IF dataset. For IF, we use AllenAI's RLVR-IFeval, whose prompts contain programmatically verifiable constraints. Each domain contributes 7.5K training examples, resulting in a balanced mixture across four domains.

Evaluation. We evaluate mathematical reasoning on AIME 2024, AIME 2025, and AMC 2023; code generation on LiveCodeBench; IF on IFBench; and Chat on WildBench-Hard and CreativeWriting v3. For AIME, AMC, and LiveCodeBench, we report avg@8. We report WildBench-Hard and CreativeWriting v3 on their normalized scales. Because the training mixture contains four equally represented domains, we use a domain-macro average as the primary aggregate metric.

Table 2: **Main results on Qwen3-8B-Base**. OSOL achieves the best domain-macro average while ranking first or second in every individual domain.

Method	Math	Code	IF	Chat	Domain Avg.
Base model	0.1316	0.1751	0.1433	0.4466	0.2241
GRPO	0.2906	0.2119	0.2733	0.6284	0.3511
GRPO+$KL_{0.001}$	0.2750	**0.2249**	**0.3000**	**0.6802**	<u>0.3700</u>
MGS	0.2948	0.0883	0.2833	0.6401	0.3266
CGPO	<u>0.3094</u>	0.1828	0.2567	0.6623	0.3528
OSOL (ours)	**0.3111**	<u>0.2173</u>	<u>0.2933</u>	<u>0.6723</u>	**0.3735**

5.2 Main Results

Tables 1 and 2 show that OSOL achieves the best domain-macro performance on both evaluated backbones. On Qwen3-30B-A3B, OSOL reaches a domain-macro average of 0.4822, outperforming MGS, the strongest baseline, by 0.0262 points or 5.7%, while achieving the best results in both mathematics and code generation. On Qwen3-8B-Base, OSOL again obtains the highest domain-macro average of 0.3735, representing a 6.4% improvement over standard GRPO, and ranks first or second in every individual domain. These results demonstrate consistent aggregate gains across two substantially different model scales and architectures.

More importantly, the relative effectiveness of existing optimization strategies varies markedly across backbones. MGS is the strongest baseline on Qwen3-30B-A3B but degrades substantially on Qwen3-8B-Base, particularly in code generation, whereas GRPO+$KL_{0.001}$ performs strongly on the 8B model but provides little improvement over standard GRPO on the 30B model. In contrast, OSOL consistently achieves the highest domain-macro performance on both backbones. This cross-backbone consistency aligns with our motivation that cross-step interference is not fully captured by within-step gradient geometry. By conditioning optimization on the realized policy drift inherited

Table 3: **Ablation results on Qwen3-8B-Base**. OSOL without history residual retains the one-step-one-lead focus structure while removing the checkpoint-history correction; the remaining variants ablate focus weighting, drift-based ranking, or token-level allocation.

Method	Math	Code	IF	Chat	Domain Avg.
OSOL without history residual	0.2996	<u>0.2114</u>	<u>0.2966</u>	0.6324	0.3600
Equal task weights	0.0615	0.0202	0.2467	0.3978	0.1815
Response-level residual	**0.3202**	0.2021	**0.3000**	<u>0.6338</u>	<u>0.3640</u>
Random residual	0.0698	0.0390	0.2400	0.3866	0.1838
OSOL (ours)	<u>0.3111</u>	**0.2173**	0.2933	**0.6723**	**0.3735**

from the preceding checkpoint, OSOL explicitly targets interference emerging across consecutive updates. We examine this mechanism directly in Section 5.4.

5.3 Ablation Analysis

We ablate OSOL to answer whether checkpoint-history correction provides gains beyond the focus-structured base update and which design choices are responsible for its effectiveness. We first remove the history residual as a whole, and then isolate focus-dependent weighting, drift-based allocation, and token-level correction. The matched OSOL without history residual retains the same focus schedule and asymmetric base weighting, but removes r_i^{OSOL} from the final coefficient in Eq. (21). Removing the history residual lowers the average from 0.3735 to 0.3600; the residual thus contributes a 3.8% relative gain beyond the focus-structured update itself.

The remaining variants examine how the correction should be constructed. Equal task weights sets $w_{\mathrm{f}} = w_{\mathrm{nf}}$ in Eq. (15) and reduces the domain-macro average to 0.1815, showing that asymmetric focus weighting is important when the correction is restricted to the designated focus domain. Random residual preserves the negative-drift candidate set in Eq. (16) but replaces the drift-ranked allocation in Eq. (17) with random weights, yielding 0.1838. This indicates that, in this configuration, preserving the candidate set alone is insufficient; the drift-based allocation also matters. Finally, Response-level residual retains the checkpoint signal but replaces token-specific allocation with a coarser response-level correction, reaching 0.3640. Restoring token-level allocation raises the average to 0.3735, consistent with the token-wise backtracking quantity in Eq. (9). Together, these results support the main components of OSOL: focus-dependent weighting, drift-informed residual allocation, and token-level correction.

5.4 Mechanism Analysis

Section 3 identifies the negative-to-positive backtracking mass $B^{-}(\boldsymbol{u}, \boldsymbol{d})$ as the quantity targeted by OSOL. We first test whether this realized cross-step quantity is associated with subsequent task degradation. Figure 3 compares three replicate-level diagnostics against held-out math damage after a subsequent chat update. The global and module-level gradient diagnostics yield Spearman correlations of -0.19 and -0.22, respectively, whereas cross-step backtracking reaches $\rho_s = 0.40$ with a 95% bootstrap CI of $[0.05, 0.69]$. The backtracking and damage measurements use disjoint probes (Appendix A.2). Paired bootstrap comparisons further give gains of $+0.59$ (95% CI $[0.05, 1.06]$) over the global diagnostic and $+0.63$ (95% CI $[0.03, 1.15]$) over the module diagnostic (Appendix B). These results indicate that realized reversal between consecutive updates is more informative of subsequent task damage than the within-step diagnostics considered here.

Realized backtracking, however, requires both the preceding footprint $\boldsymbol{u}$ and the current footprint $\boldsymbol{d}$, and is observable only after the current update has been applied. The online controller cannot use $\boldsymbol{d}$ at decision time, so its cross-step history signal is limited to the preceding footprint $\boldsymbol{u}$. We next test whether $\boldsymbol{u}$ alone can identify negative-drift tokens that are likely to rebound. Across the eight primary seeds, checkpoint-footprint ranking achieves 0.567 ROC-AUC and 0.627 average precision within the negative-drift candidate set, compared with 0.498 and 0.579 for the Hessian proxy, and improves top-5% precision from 0.551 to 0.661. Thus, the preceding drift provides a useful pre-update ranking of future rebound risk.

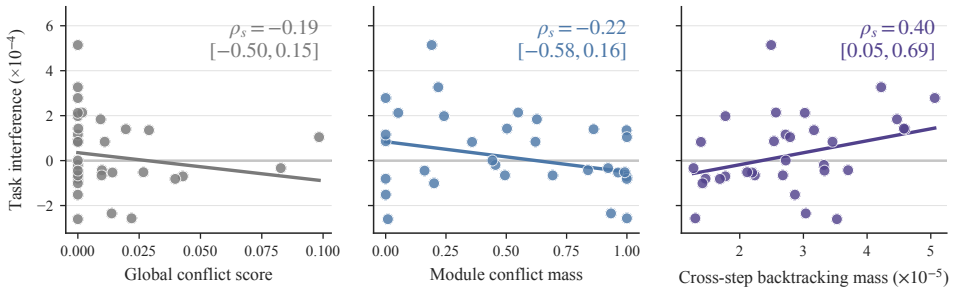

Figure 3: **Replicate-level association between interference diagnostics and held-out task damage.** Cross-step backtracking is positively associated with subsequent damage, whereas the global and module gradient diagnostics are not. Backtracking and damage are measured on disjoint probes.

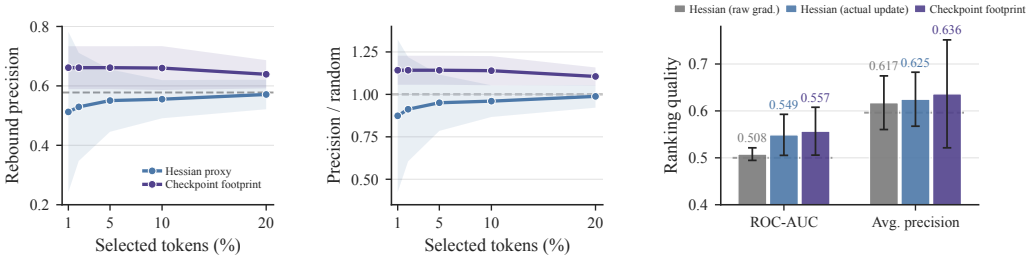

Figure 4: **The preceding checkpoint footprint better identifies tokens that rebound in the next update.** Its advantage persists across selection fractions and remains after strengthening the Hessian proxy with the actual optimizer displacement instead of the raw gradient direction.

Figure 4 tests the robustness of this token-level signal. Checkpoint-footprint ranking remains stronger across selection fractions from 1% to 20%, with the largest advantage among the highest-ranked tokens, consistent with the drift-ranked allocation in Eq. (17). We further test whether the Hessian proxy is disadvantaged simply because it uses the raw-gradient direction. On a three-seed subset, replacing this direction with the actual optimizer displacement raises its ROC-AUC from 0.508 to 0.549, while checkpoint-footprint ranking remains higher at 0.557. The checkpoint signal's advantage is therefore not explained solely by update-direction mismatch. Together, these results link the theory to the online algorithm: B^- captures realized cross-step reversal, while $\boldsymbol{u}$ provides the pre-update signal used to identify coordinates at risk of future rebound.

6 Conclusion

This work shows that cross-domain interference in multi-domain RL can emerge across consecutive updates and remain hidden from within-step gradient diagnostics. We characterize this sequential interference through token log-probability footprints, where opposing movements reveal cross-step output backtracking, and propose OSOL to use the preceding checkpoint footprint as an online signal for correcting rebound-prone tokens within standard GRPO. Across Qwen3-8B-Base and Qwen3-30B-A3B, OSOL achieves the best domain-macro performance among the compared methods; on Qwen3-30B-A3B, it improves the domain-macro average by 5.7% over the strongest baseline and yields a 17.4% gain in code generation. Controlled studies further show that cross-step backtracking is more informative of subsequent task damage than within-step gradient diagnostics, while the preceding checkpoint footprint identifies future rebound more effectively than Hessian-based proxies. More broadly, our results suggest that the realized optimization trajectory itself can provide a useful signal for controlling multi-domain RL, offering an alternative to estimating cross-domain interactions solely from within-step gradient geometry. Future work may extend this perspective to longer update histories, adaptive focus scheduling, and broader reasoning and agentic training settings.

AI USE STATEMENT

We used GPT-5.6 to assist with language editing and improve the manuscript's clarity. The authors take full responsibility for the final text and for verifying the accuracy of all statements and claims.

REPRODUCIBILITY STATEMENT

Appendix A specifies the protocols of the two controlled diagnostic studies, including models, data, one-step optimization settings, scoring rules, and statistical procedures; Appendix B reports the corresponding uncertainty, stability, and sensitivity analyses. Appendices C–E state the local assumptions and give complete proofs of all results in Section 3. Appendix F details the residual construction, including degenerate cases, and Table 4 lists the main training, generation, and OSOL-specific hyperparameters. Code is available at `https://anonymous.4open.science/r/anonymous-osol-83DF/`.

REFERENCES

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, et al. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*, 2025.

Min Cai, Yu Liang, Longzheng Wang, Yan Wang, Yueyang Zhang, Long Xia, Zhiyuan Sun, Xi Ye, and Daiting Shi. Advancing general-purpose reasoning models with modular gradient surgery. *arXiv preprint arXiv:2602.02301*, 2026.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Jorge Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Nilabjo Dey, Yonghao Zhuang, Yuheng Zha, et al. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. *Advances in Neural Information Processing Systems*, 38, 2026.

Zi Gong, Hang Yu, Cong Liao, Bingchang Liu, Chaoyu Chen, and Jianguo Li. Coba: Convergence balancer for multitask finetuning of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8063–8077, 2024.

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

Zhexin Hu, Li Wang, Xiaohan Wang, Jiajun Chai, Xiaojun Guo, Wei Lin, and Guojun Yin. Ziprl: Adaptive multi-turn context compression with hindsight response replay. *arXiv preprint arXiv:2605.28069*, 2026.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32, 2019.

Yu Li, Zhuoshi Pan, Honglin Lin, Mengyuan Sun, Conghui He, and Lijun Wu. Can one domain help others? a data-centric study on multi-domain reasoning via reinforcement learning. *arXiv preprint arXiv:2507.17512*, 2025.

Xize Liang, Lin Yang, Jie Wang, Rui Liu, Yang Lu, Jinliang Zeng, Hanzhu Chen, Dong Li, and Jianye Hao. Boosting multi-domain reasoning of llms via curvature-guided policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. Famo: Fast adaptive multitask optimization. *Advances in Neural Information Processing Systems*, 36:57226–57243, 2023.

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. *arXiv preprint arXiv:2602.08499*, 2026.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Shyam Sundhar Ramesh, Xiaotong Ji, Matthieu Zimmer, Sangwoong Yoon, Zhiyong Wang, Haitham Bou Ammar, Aurelien Lucchi, and Ilija Bogunovic. Multi-task grpo: Reliable llm reasoning across tasks. *arXiv preprint arXiv:2602.05547*, 2026.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Runzhe Wu, Ankur Samanta, Ayush Jain, Scott Fujimoto, Jeongyeol Kwon, Ben Kretzu, Youliang Yu, Kaveh Hassani, Boris Vidolov, and Yonathan Efroni. Imbalanced gradients in rl post-training of multi-task llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3137–3150, 2026.

Lei Yang, Siyu Ding, and Deyi Xiong. A local perturbation theory for cross-domain interference and recovery in multi-domain rl. *arXiv preprint arXiv:2606.02398*, 2026a.

Wang Yang, Shouren Wang, Chaoda Song, Chuang Ma, Xinpeng Li, Nengbo Wang, Kaixiong Zhou, Vipin Chaudhary, and Xiaotian Han. When domains interact: Asymmetric and order-sensitive cross-domain effects in reinforcement learning for reasoning. *arXiv preprint arXiv:2602.01365*, 2026b.

Yongjin Yang, Jiarui Liu, Yinghui He, Lezhen Zhang, Bernhard Schölkopf, and Zhijing Jin. Transferability for general reasoning: An automated curriculum for multi-domain rlvr. *arXiv preprint arXiv:2606.25178*, 2026c.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

A CONTROLLED DIAGNOSTIC STUDY PROTOCOLS

This section provides the complete protocols for the two controlled studies used to motivate and analyze OSOL. Both studies instantiate adjacent updates explicitly, preserve the optimizer state across the two steps, and evaluate every diagnostic against quantities measured from the realized checkpoint trajectory. The first study evaluates which replicate-level signal best explains downstream cross-domain interference. The second evaluates which token-level signal best identifies tokens whose preceding output movement the next update reverses.

A.1 COMMON MODEL, DATA, AND ONE-STEP OPTIMIZATION

Model and data. We use Qwen3-4B as the policy in both studies. Math prompts are sampled from the math portion of the same filtered NVIDIA Nemotron Post-Training Dataset v2 mixture used for the main experiments, and chat prompts are sampled from the MGS-provided WildChat-IF subset. The task verifier scores math generations, and Skywork-Reward-Llama-3.1-8B-v0.2 scores chat generations. For each prompt, we sample four responses with temperature 1.0 and nucleus threshold $p = 0.95$. Prompts are truncated to at most $1{,}024$ tokens and responses to at most 512 tokens. We disable explicit thinking-mode prompting so all diagnostics use the same response format.

Advantages and updates. Within each four-response group, we compute the stabilized GRPO trajectory advantage $A_j^{\mathrm{g}} = (R_j - \bar{R})/\sqrt{\sigma_R^2 + \epsilon_{\mathrm{adv}}}$, with the same $\epsilon_{\mathrm{adv}} > 0$ as in Eq. (14). For the controlled diagnostics only, an equal-reward group is assigned a unit coefficient, yielding a deterministic nonzero one-step update while preserving the sampled response distribution. Chat groups use reward-model scores and have nonzero within-group variation in the analyzed runs. Each replicate performs two consecutive AdamW steps with learning rate 10^{-6}, weight decay 0.1, $(\beta_1, \beta_2) = (0.9, 0.999)$, numerical constant 10^{-8}, and gradient-norm clipping at 1.0. Computation uses bfloat16, and log probabilities are accumulated over response tokens with padding coordinates masked as appropriate.

Controlled output-facing subspace. The repeated gradient, optimizer, and finite-difference evaluations are carried out in an output-facing parameter subspace consisting of the final transformer block (layer 35 of Qwen3-4B). This subspace contains $100{,}930{,}816$ parameters, or 2.51% of Qwen3-4B, and is shared across the diagnostics compared in the primary studies. The coverage control in Appendix A.4 expands this subspace to the final four and final eight transformer blocks. This design permits exact replicate-level replay of gradients, optimizer displacements, and token log probabilities while keeping the policy, sampled trajectories, rewards, and evaluation targets matched.

A.2 STUDY I: SINGLE-CHECKPOINT DIAGNOSTICS AND REALIZED TASK INTERFERENCE

Replicate construction. We construct 32 independent replicates from a common initial checkpoint $\boldsymbol{\theta}_0$. Randomness across replicates comes from prompt selection and response sampling; the initial policy and all optimization hyperparameters are fixed. Each replicate uses two math-update prompts and two chat-update prompts, with four responses per prompt. In addition, it samples two math prompts for the token-motion probe and two disjoint math prompts for the damage probe. The update, motion-probe, and damage-probe prompt sets are mutually disjoint within a replicate.

Starting from $\boldsymbol{\theta}_0$, we first apply the math update and then the chat update:

$$\boldsymbol{\theta}_0 \xrightarrow{\text{math update}} \boldsymbol{\theta}_1 \xrightarrow{\text{chat update}} \boldsymbol{\theta}_2. \tag{23}$$

Math responses and both math probe sets are sampled at $\boldsymbol{\theta}_0$. After the math step, chat responses are sampled from $\boldsymbol{\theta}_1$, scored by the reward model, and used for the $\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2$ update. To obtain a single-checkpoint gradient comparison without changing the realized chat trajectories, the fixed $\boldsymbol{\theta}_1$ chat responses, rewards, and advantages are replayed at $\boldsymbol{\theta}_0$ when computing the chat gradient. We retain the same AdamW optimizer instance across the math and chat steps, matching the stateful sequential-update dynamics.

Independent interference target. Let $\mathcal{P}_{\mathrm{dmg}}$ denote the held-out math damage probe. We measure the effect of the chat update by the change in fixed-response token negative log-likelihood:

$$\mathrm{Damage}_{\mathrm{math}\leftarrow\mathrm{chat}} := \mathcal{L}_{\mathrm{NLL}}(\boldsymbol{\theta}_2; \mathcal{P}_{\mathrm{dmg}}) - \mathcal{L}_{\mathrm{NLL}}(\boldsymbol{\theta}_1; \mathcal{P}_{\mathrm{dmg}}). \tag{24}$$

A positive value indicates that the chat step increases math-probe NLL. Because $\mathcal{P}_{\mathrm{dmg}}$ is disjoint from the update and motion probes, the dependent variable is not reused in any diagnostic score.

Single-checkpoint global and module diagnostics. Using the fixed update batches, define the descent directions at $\boldsymbol{\theta}_0$ as

$$\boldsymbol{v}_{\mathrm{m}} := -\nabla \mathcal{L}_{\mathrm{m}}(\boldsymbol{\theta}_0), \qquad \boldsymbol{v}_{\mathrm{c}} := -\nabla \mathcal{L}_{\mathrm{c}}(\boldsymbol{\theta}_0). \tag{25}$$

The global first-order diagnostic is

$$s_{\mathrm{global}} = \left[-\frac{\langle \boldsymbol{v}_{\mathrm{m}}, \boldsymbol{v}_{\mathrm{c}} \rangle}{\|\boldsymbol{v}_{\mathrm{m}}\|_2 \|\boldsymbol{v}_{\mathrm{c}}\|_2} \right]_+ . \tag{26}$$

For the module diagnostic, partition the output-facing subspace into attention projections, MLP projections, and normalization parameters. If $\boldsymbol{v}_{\mathrm{m}}^{(g)}$ and $\boldsymbol{v}_{\mathrm{c}}^{(g)}$ are the directions in group g, define

$$s_{\mathrm{module}} = \frac{\sum_g [-\langle \boldsymbol{v}_{\mathrm{m}}^{(g)}, \boldsymbol{v}_{\mathrm{c}}^{(g)} \rangle]_+}{\sum_g |\langle \boldsymbol{v}_{\mathrm{m}}^{(g)}, \boldsymbol{v}_{\mathrm{c}}^{(g)} \rangle|}. \tag{27}$$

This normalized negative dot-product mass retains opposing module contributions that cancel in the aggregate inner product.

Cross-step backtracking diagnostic. On the disjoint math motion probe $\mathcal{P}_{\mathrm{mot}}$, we evaluate the same realized response tokens at all three checkpoints. For token i, let

$$u_i = \log \pi_{\boldsymbol{\theta}_1}(a_i \mid x_i) - \log \pi_{\boldsymbol{\theta}_0}(a_i \mid x_i), \qquad d_i = \log \pi_{\boldsymbol{\theta}_2}(a_i \mid x_i) - \log \pi_{\boldsymbol{\theta}_1}(a_i \mid x_i). \tag{28}$$

The backtracking mass used in Figure 3 is

$$s_{\mathrm{rev}} = \frac{1}{|\mathcal{R}_{\mathrm{mot}}|} \sum_{i \in \mathcal{R}_{\mathrm{mot}}} \mathbf{1}\{u_i < 0\} \, [-u_i d_i]_+ = \frac{1}{|\mathcal{R}_{\mathrm{mot}}|} \sum_{i \in \mathcal{R}_{\mathrm{mot}}} (-u_i)[d_i]_+ \, \mathbf{1}\{u_i < 0\}, \tag{29}$$

where $\mathcal{R}_{\mathrm{mot}}$ contains response-token coordinates. The diagnostic therefore records the mass of positive rebound following a negative preceding movement, using the realized optimizer steps rather than a virtual parameter perturbation.

Statistical analysis. Each replicate contributes one value of $\mathrm{Damage}_{\mathrm{math} \leftarrow \mathrm{chat}}$ and one value for each diagnostic. We use Spearman rank correlation as the primary association statistic and display a Theil–Sen trend in the scatter plots. We obtain confidence intervals from $50{,}000$ paired replicate bootstrap resamples. We use the same resampled replicate indices when comparing two diagnostics, so the reported correlation gains are paired differences rather than differences between independently estimated intervals. We additionally recompute each correlation after removing every replicate once. The resulting leave-one-replicate-out ranges are reported in Figure 8.

Observed diagnostic ordering. The global and module diagnostics have correlations of -0.19 and -0.22 with downstream interference, whereas backtracking mass reaches $\rho_s = 0.40$ with a 95% bootstrap interval of $[0.05, 0.69]$. Its paired correlation gains are $+0.59$ over the global diagnostic and $+0.63$ over the module diagnostic. The sign and ordering are preserved under every leave-one-replicate-out recomputation. Figure 6 further decomposes the same replicates and shows how positive and negative module contributions disappear under global aggregation.

A.3 Study II: token-level rebound identification

Causal triplets and token table. The second study uses eight independent seeds and the same Qwen3-4B one-step configuration. Each seed samples two math prompts and two chat prompts with four responses per prompt. The math step produces $\boldsymbol{\theta}_1$ from $\boldsymbol{\theta}_0$. Chat responses are then sampled from $\boldsymbol{\theta}_1$, scored by the reward model, and used both as the token probe and as the update batch that produces $\boldsymbol{\theta}_2$. Therefore, u_i and d_i in Eq. (28) are evaluated on the exact token coordinates responsible for the second update.

We retain chat response tokens with nonzero trajectory-level advantage:

$$\mathcal{E} = \{i : \ i \text{ is a response token}, \ A_i \neq 0\}. \tag{30}$$

The checkpoint-history candidate set and realized rebound label are

$$\mathcal{C} = \{i \in \mathcal{E} : u_i < 0\}, \qquad y_i = \mathbf{1}\{u_i < 0,\ d_i > 0\}, \qquad b_i^{\mathrm{reb}} = (-u_i)[d_i]_+. \tag{31}$$

The primary comparison is conditioned on $\mathcal{C}$: it asks which negative-drift tokens are most likely to rebound under the subsequent unregulated update. Across the eight seeds, 22.6% of eligible tokens enter $\mathcal{C}$, and the mean rebound prevalence within $\mathcal{C}$ is 57.8%.

Checkpoint-footprint score. Within $\mathcal{C}$, the raw checkpoint score is $[-u_i]_+$. The rank-based score used by OSOL sorts candidate drifts in ascending order and assigns

$$p_i = \frac{\mathrm{rank}_0(u_i) + 1/2}{|\mathcal{C}|}, \qquad s_i^{\mathrm{ckpt}} = 1 - \frac{\mathrm{clip}(p_i, 0.1, 1) - 0.1}{0.9}, \tag{32}$$

with score zero outside $\mathcal{C}$. More negative drift gets a higher score, while the strongest decile gets the maximum score. This is the same ordering and quantile profile used by the online controller before batchwise scale normalization.

Finite-difference Hessian proxy. We compare checkpoint history with a token-level curvature proxy constructed on the same fixed chat trajectories, rewards, and advantages. Let $\mathcal{L}_{\mathrm{c}}^{\mathrm{loc}}$ be the unclipped local policy-gradient surrogate on this fixed chat batch, and let

$$\boldsymbol{v}_{\mathrm{m}} := -\nabla \mathcal{L}_{\mathrm{m}}(\boldsymbol{\theta}_0), \qquad \boldsymbol{v}_{\mathrm{c}} := -\nabla \mathcal{L}_{\mathrm{c}}^{\mathrm{loc}}(\boldsymbol{\theta}_0). \tag{33}$$

For selected parameters $\boldsymbol{\theta}_{\mathcal{S}}$, choose

$$\alpha = \epsilon \frac{\|\boldsymbol{\theta}_{\mathcal{S}}\|_2}{\|\boldsymbol{v}_{\mathrm{m}}\|_2} \tag{34}$$

and estimate the directional Hessian-vector product by the centered difference

$$\boldsymbol{h}_{\mathrm{HVP}} \approx \frac{\nabla \mathcal{L}_{\mathrm{c}}^{\mathrm{loc}}(\boldsymbol{\theta}_0 + \alpha \boldsymbol{v}_{\mathrm{m}}) - \nabla \mathcal{L}_{\mathrm{c}}^{\mathrm{loc}}(\boldsymbol{\theta}_0 - \alpha \boldsymbol{v}_{\mathrm{m}})}{2\alpha}. \tag{35}$$

The main sweep uses $\epsilon \in \{10^{-5}, 3{\times}10^{-5}, 10^{-4}, 3{\times}10^{-4}\}$; the primary score uses the pre-specified setting $\epsilon = 10^{-4}$. Define

$$\boldsymbol{\theta}_{\mathrm{m}} = \boldsymbol{\theta}_0 + \alpha \boldsymbol{v}_{\mathrm{m}}, \qquad \boldsymbol{\theta}_{\mathrm{pred}} = \boldsymbol{\theta}_{\mathrm{m}} + \alpha(\boldsymbol{v}_{\mathrm{c}} - \alpha \boldsymbol{h}_{\mathrm{HVP}}). \tag{36}$$

For probe token i, write $\ell_i(\boldsymbol{\theta}) := \log \pi_{\boldsymbol{\theta}}(a_i \mid x_i)$. The virtual math displacement and curvature-adjusted predicted chat displacement are

$$\Delta_i^{\mathrm{m}} = \ell_i(\boldsymbol{\theta}_{\mathrm{m}}) - \ell_i(\boldsymbol{\theta}_0), \qquad \Delta_i^{\mathrm{pred}} = \ell_i(\boldsymbol{\theta}_{\mathrm{pred}}) - \ell_i(\boldsymbol{\theta}_{\mathrm{m}}), \tag{37}$$

and the Hessian ranking score is

$$s_i^{\mathrm{H}} = [-\Delta_i^{\mathrm{m}} \Delta_i^{\mathrm{pred}}]_+. \tag{38}$$

We compute all virtual log probabilities on the same tokenized probe batch. We also record the induced forward policy KL for each virtual displacement, providing an output-space check on the numerical perturbation scale.

Ranking metrics and uncertainty. For every seed, we compute receiver operating characteristic area under the curve (ROC-AUC), average precision, and precision among the top-ranked 1%, 2%, 5%, 10%, and 20% of candidates. Enrichment is precision divided by the within-candidate rebound prevalence. Average precision is evaluated at unique score thresholds, and when a top-k boundary intersects tied scores, we report expected precision under uniform tie-breaking. We compute each metric independently for each seed, then report the seed mean with a two-sided 95% Student-t interval. Paired gains use the per-seed difference between checkpoint-footprint and Hessian scores before aggregation.

Primary comparison. Within negative-drift candidates, checkpoint-footprint ranking obtains 0.567 ROC-AUC, 0.627 average precision, and 0.661 top-5% precision, compared with 0.498, 0.579, and 0.551 for the Hessian proxy. The paired gains in ROC-AUC and average precision are 0.069 and 0.048, respectively. Across selected candidate fractions, checkpoint ranking remains above the Hessian proxy in both precision and enrichment, with the clearest separation when only a small fraction of candidates is selected by OSOL.

A.4 HESSIAN CONTROLS AND COMPUTATIONAL MEASUREMENTS

Realized-update direction. To separate curvature approximation from direction mismatch, we repeat the Hessian construction on seeds $\{11, 23, 37\}$ after replacing $\boldsymbol{v}_{\mathrm{m}}$ with the selected-parameter optimizer displacement $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0$. The finite-difference procedure and token target remain unchanged. This increases Hessian ROC-AUC from 0.508 to 0.549 and average precision from 0.617 to 0.625; checkpoint footprint remains highest at 0.557 and 0.636, respectively.

Parameter-coverage sweep. We repeat the curvature proxy on the final one, four, or eight transformer blocks, corresponding to 2.51%, 10.04%, and 20.07% of model parameters. Each configuration uses the same three seeds and finite-difference scales $\{3\times10^{-5}, 10^{-4}\}$. We measure peak GPU memory as the maximum CUDA allocation during the complete diagnostic run. Mean peak memory rises from 26.6 to 33.8 and 43.5 GB, while Hessian ROC-AUC changes from 0.508 to 0.511 and 0.534. Checkpoint-footprint ROC-AUC on the matching trajectories is 0.557, 0.577, and 0.583.

Finite-difference scale sweep. Finally, we retain all eight primary seeds and recompute the finite-difference diagnostic for $\epsilon \in \{10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$. For each scale, we record the HVP norm and the forward policy KL induced by the virtual math step, the displacement induced by the negative HVP term, and the predicted next step. These measurements produce Figure 9 and quantify how the numerical probe scale changes the output-space displacement being evaluated.

B Additional Diagnostic Evidence

This section expands the controlled analyses in Appendix A. The primary Study I comparison between realized cross-step backtracking and held-out task damage is reported in Section 5.4; here we provide module-level decomposition and robustness controls for that result. Study II evaluates pre-update token-rebound ranking; its primary eight-seed comparison is shown in Figure 5, followed by numerical controls for the finite-difference Hessian proxy.

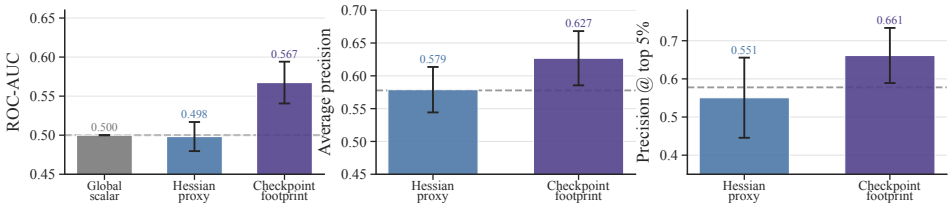

Figure 5: **Rebound identification within negative-drift candidates**. The checkpoint footprint outperforms the Hessian proxy in ROC-AUC, average precision, and top-5% precision; the non-token-resolved global scalar is included as a 0.5 ROC-AUC reference. Error bars show 95% confidence intervals.

Module-level decomposition exposes cancellation. Figure 6 decomposes the single-checkpoint gradient interaction across attention projections, MLP projections, and normalization parameters. Opposing module contributions can cancel in the global cosine, but retaining this finer decomposition does not recover a positive ordering of subsequent damage: the module conflict mass has $\rho_s = -0.22$ in the controlled Study I comparison. The figure therefore isolates one failure mode of global aggregation while leaving the trajectory-level comparison to Section 5.4.

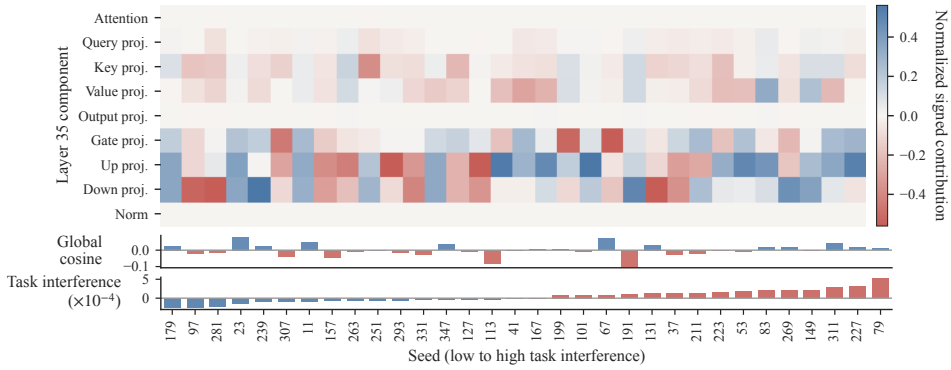

Figure 6: **Layer-35 signed gradient contributions expose cancellation hidden by the global cosine.** Columns are seeds ordered by observed task interference; opposing module contributions can cancel under global aggregation.

Paired uncertainty supports the main-text comparison. Figure 7 reports the bootstrap intervals of the three Spearman correlations and their paired differences. The intervals for the global and module diagnostics include zero, whereas the interval for backtracking mass lies on the positive side. Using the same bootstrap resamples across diagnostics, backtracking improves correlation by $+0.59$ over the global score (95% CI $[0.05, 1.06]$) and by $+0.63$ over the module score (95% CI $[0.03, 1.15]$). These paired comparisons account for replicate-to-replicate variation and provide the uncertainty analysis for the result reported in Section 5.4.

The diagnostic ordering is not driven by a single replicate. The controlled study contains 32 replicates, so we additionally assess sensitivity to individual runs rather than relying on the aggregate correlation alone. Figure 8 recomputes each Spearman correlation after removing one replicate

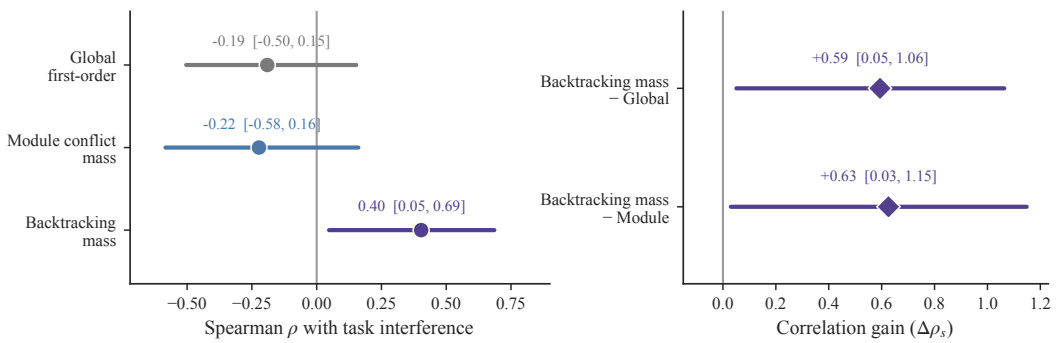

Figure 7: **Uncertainty and paired correlation gains for the three Study I diagnostics**. Left: bootstrap estimates of Spearman correlation with task damage. Right: paired gains from replacing the global or module gradient diagnostic with cross-step backtracking mass. Both paired intervals remain positive.

at a time. Across all 32 leave-one-out subsets, the global score remains negative in $[-0.26, -0.15]$, the module conflict mass remains negative in $[-0.31, -0.16]$, and backtracking mass remains positive in $[0.35, 0.49]$. Thus every deletion preserves both the sign pattern and the ordering of the three diagnostics. This analysis is not a substitute for an independent replication study, but it rules out the simpler explanation that one unusually influential replicate drives the positive backtracking correlation, or the gap to the within-step baselines.

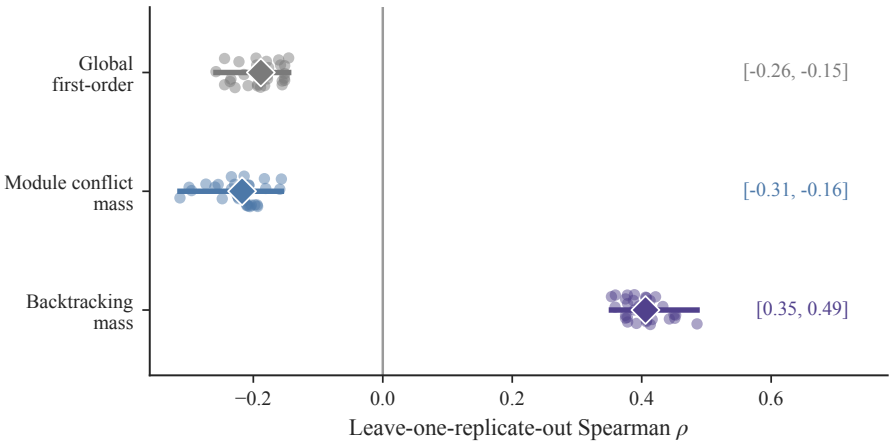

Figure 8: **Leave-one-replicate stability of the Study I correlations**. Each point recomputes Spearman ρ after removing one replicate; horizontal ranges summarize all 32 deletions. The sign and ordering of the diagnostics are preserved throughout.

Finite-difference curvature introduces a consequential probe-scale choice. Figure 9 examines the finite-difference scale ϵ used to construct the Hessian-vector-product proxy in Study II. The left panel shows that the estimated HVP norm changes markedly at the smallest probe scale, then flattens as ϵ increases. The right panel shows the corresponding output-space effect: the policy KL induced by the virtual focus step, the Hessian reaction, and the resulting predicted next step all increase by orders of magnitude over the same sweep. Hence ϵ affects not only the numerical value of the curvature estimate but also the size of the virtual policy displacement on which the token-ranking proxy is evaluated. This introduces an additional sensitivity for online use: changing the probe scale changes the effective output-space regime of the diagnostic even when the underlying checkpoint and trajectories are fixed. The checkpoint-footprint signal used by OSOL does not require this additional finite-difference probe; it measures the log-probability change between realized adjacent checkpoints directly. Figure 9 therefore complements the accuracy and memory controls in Ap-

pendix A.4: the limitation of the Hessian route is not only computational cost, but also the need to choose a numerical perturbation scale whose effect propagates into the policy-space quantity being ranked.

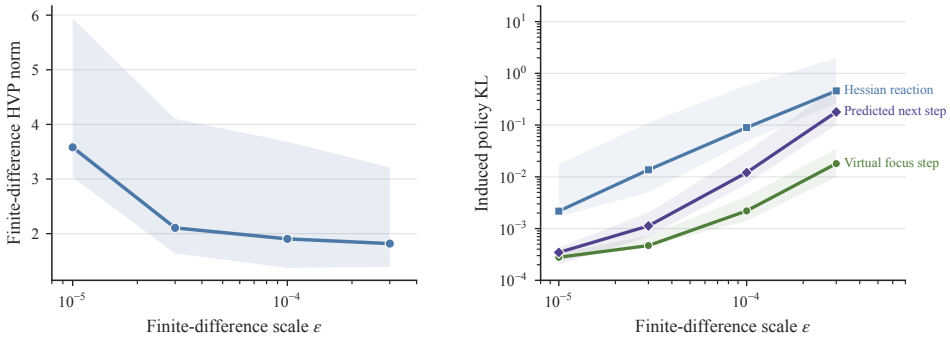

Figure 9: **Sensitivity of the finite-difference Hessian proxy to the probe scale** ϵ. Left: the estimated HVP norm varies with ϵ. Right: the induced policy KL of the virtual focus step, Hessian reaction, and predicted next step changes by orders of magnitude, showing that the probe scale also changes the output-space displacement being evaluated.

Taken together, these controls sharpen the mechanism evidence without changing the role assigned to each study in the main text. Study I shows that realized cross-step backtracking is a stronger post-update diagnostic of task damage than within-step gradient scores. Study II shows that finite-difference curvature introduces an additional probe-scale choice whose effect propagates into the output-space quantity used for token ranking. The online controller instead uses the preceding checkpoint footprint, which is already available before the next update and whose rebound-ranking ability is evaluated in Section 5.4. The results are therefore complementary: realized two-step backtracking characterizes the interaction that occurred, while the preceding footprint provides the pre-update information used to determine the intervention set.

C Output-Space Geometry

C.1 Notation and local regularity

Fix iteration k and the eligible focus coordinates $\mathcal{E}_k = \{(x_i, a_i)\}_{i=1}^N$. For an arbitrary action a at context x_i, define

$$\ell_{i,a}(\boldsymbol{\theta}) := \log \pi_{\boldsymbol{\theta}}(a \mid x_i), \qquad \boldsymbol{\varphi}_{i,a} := \nabla \ell_{i,a}(\boldsymbol{\theta}_k). \tag{39}$$

For the realized action a_i, abbreviate

$$\ell_i(\boldsymbol{\theta}) := \ell_{i,a_i}(\boldsymbol{\theta}), \qquad \boldsymbol{\varphi}_i := \boldsymbol{\varphi}_{i,a_i}, \qquad \boldsymbol{\ell}_k(\boldsymbol{\theta}) := [\ell_1(\boldsymbol{\theta}), \dots, \ell_N(\boldsymbol{\theta})]^\top. \tag{40}$$

Collect the realized scores into

$$\boldsymbol{\Phi}_k := [\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_N] \in \mathbb{R}^{P \times N}, \qquad \boldsymbol{K}_k := \boldsymbol{\Phi}_k^\top \boldsymbol{\Phi}_k \in \mathbb{R}^{N \times N}. \tag{41}$$

Then

$$\boldsymbol{F}_k = \frac{1}{N} \boldsymbol{\Phi}_k \boldsymbol{\Phi}_k^\top, \qquad \overline{\boldsymbol{F}}_k = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{a \sim \pi_{\boldsymbol{\theta}_k}(\cdot \mid x_i)} [\boldsymbol{\varphi}_{i,a} \boldsymbol{\varphi}_{i,a}^\top]. \tag{42}$$

We use the following local conditions, standard for Taylor and trust-region analyses (Nesterov, 2013; Nocedal & Wright, 2006).

1. The policy support is fixed in a neighborhood $\mathcal{N}_k$ of $\boldsymbol{\theta}_k$, and $\ell_{i,a}$ is three times continuously differentiable on $\mathcal{N}_k$.
2. There are constants $G, L, M < \infty$ such that, for every relevant (i, a) and $\boldsymbol{\theta} \in \mathcal{N}_k$,
$$\|\nabla \ell_{i,a}(\boldsymbol{\theta})\| \le G, \qquad \|\nabla^2 \ell_{i,a}(\boldsymbol{\theta})\|_{\mathrm{op}} \le L, \qquad \|\nabla^3 \ell_{i,a}(\boldsymbol{\theta})\|_{\mathrm{op}} \le M. \tag{43}$$
3. The line segments joining $\boldsymbol{\theta}_{k-1}$, $\boldsymbol{\theta}_k$, $\boldsymbol{\theta}_{k+1}^0$, and the locally corrected endpoint remain in $\mathcal{N}_k$.
4. The adjacent displacements satisfy
$$\|\boldsymbol{\delta}_{k-1}\| \le a\eta, \qquad \|\boldsymbol{\delta}_k^0\| \le b\eta, \qquad \|\boldsymbol{\delta}_k^r\| \le c_r \eta \tag{44}$$
for constants a, b, c_r independent of η.

The checkpoint identity in Appendix D is deterministic for the realized token coordinates and does not require independence across them. Sampling assumptions enter only when interpreting $\boldsymbol{F}_k$ as an estimator of $\overline{\boldsymbol{F}}_k$.

C.2 Relation to objective-Hessian perturbation analyses

Prior local-perturbation work studies damage to an earlier task objective after a later domain update. Around an approximately stationary earlier-task checkpoint, Taylor expansion makes the objective-Hessian quadratic term dominant (Yang et al., 2026a); curvature-guided multi-domain optimization similarly motivates explicit or approximate curvature information (Liang et al., 2026). These results explain why gradient cosine evaluated at a single checkpoint can miss sequential interference, but their primary object is the self-quadratic sensitivity of one displacement under an objective Hessian.

Our analysis changes both the scalar function and the interaction being studied. We expand the task-conditioned output KL in Eq. (2), whose exact Hessian at the anchor is the Fisher matrix, and then evaluate the mixed output interaction between the preceding realized displacement and the current base displacement. Consequently, Eq. (6) is a bilinear second-order term in output geometry, not an identification of the complete objective Hessian with the Fisher. This output-space choice follows natural-gradient and trust-region policy optimization (Amari, 1998; Kakade, 2001; Schulman et al., 2015; Martens, 2020), while making the interaction observable through checkpoint log-probabilities.

C.3 Proof that task-conditioned output KL has Fisher Hessian

For one context x_i, write

$$D_i(\boldsymbol{\delta}) := D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}_k}(\cdot \mid x_i) \,\|\, \pi_{\boldsymbol{\theta}_k+\boldsymbol{\delta}}(\cdot \mid x_i)). \tag{45}$$

Let $p_i(a) := \pi_{\boldsymbol{\theta}_k}(a \mid x_i)$. Since the first distribution is fixed with respect to $\boldsymbol{\delta}$,

$$D_i(\boldsymbol{\delta}) = \sum_a p_i(a) \log p_i(a) - \sum_a p_i(a) \ell_{i,a}(\boldsymbol{\theta}_k + \boldsymbol{\delta}). \tag{46}$$

At $\boldsymbol{\delta} = \mathbf{0}$, the two distributions coincide, so

$$D_i(\mathbf{0}) = 0. \tag{47}$$

Differentiating Eq. (46),

$$\nabla_{\boldsymbol{\delta}} D_i(\boldsymbol{\delta}) = -\sum_a p_i(a) \nabla \ell_{i,a}(\boldsymbol{\theta}_k + \boldsymbol{\delta}). \tag{48}$$

At the anchor,

$$\sum_a p_i(a) \boldsymbol{\varphi}_{i,a} = \sum_a \pi_{\boldsymbol{\theta}_k}(a \mid x_i) \nabla \log \pi_{\boldsymbol{\theta}_k}(a \mid x_i) = \sum_a \nabla \pi_{\boldsymbol{\theta}_k}(a \mid x_i) = \nabla 1 = \mathbf{0}. \tag{49}$$

Therefore

$$\nabla D_i(\mathbf{0}) = \mathbf{0}. \tag{50}$$

Differentiate the normalization identity once more. For each action,

$$\nabla^2 \pi_{\boldsymbol{\theta}}(a \mid x_i) = \pi_{\boldsymbol{\theta}}(a \mid x_i) \left[\nabla^2 \log \pi_{\boldsymbol{\theta}}(a \mid x_i) + \nabla \log \pi_{\boldsymbol{\theta}}(a \mid x_i) \nabla \log \pi_{\boldsymbol{\theta}}(a \mid x_i)^\top \right]. \tag{51}$$

Summing over a gives zero because $\sum_a \pi_{\boldsymbol{\theta}}(a \mid x_i) = 1$. Evaluated at $\boldsymbol{\theta}_k$,

$$-\sum_a p_i(a) \nabla^2 \ell_{i,a}(\boldsymbol{\theta}_k) = \sum_a p_i(a) \boldsymbol{\varphi}_{i,a} \boldsymbol{\varphi}_{i,a}^\top. \tag{52}$$

On the other hand, differentiating Eq. (48) gives

$$\nabla^2 D_i(\mathbf{0}) = -\sum_a p_i(a) \nabla^2 \ell_{i,a}(\boldsymbol{\theta}_k). \tag{53}$$

Combining Eqs. (52) and (53),

$$\nabla^2 D_i(\mathbf{0}) = \mathbb{E}_{a \sim \pi_{\boldsymbol{\theta}_k}(\cdot \mid x_i)} [\boldsymbol{\varphi}_{i,a} \boldsymbol{\varphi}_{i,a}^\top]. \tag{54}$$

Averaging over i gives $\nabla^2 \mathcal{D}_k(\mathbf{0}) = \overline{\boldsymbol{F}}_k$, the standard Fisher–KL identity (Amari, 1998; Martens, 2020).

C.4 Explicit KL Taylor remainder and mixed derivative

Apply multivariate Taylor expansion to $\mathcal{D}_k$ at $\mathbf{0}$. Since its value and gradient vanish,

$$\mathcal{D}_k(\boldsymbol{\delta}) = \frac{1}{2} \boldsymbol{\delta}^\top \nabla^2 \mathcal{D}_k(\mathbf{0}) \boldsymbol{\delta} + R_{\mathrm{KL}}(\boldsymbol{\delta}). \tag{55}$$

Under the third-derivative bound in Eq. (43), the integral or Lagrange remainder satisfies

$$|R_{\mathrm{KL}}(\boldsymbol{\delta})| \le \frac{M_{\mathcal{D}}}{6} \|\boldsymbol{\delta}\|^3 \tag{56}$$

for a local constant $M_{\mathcal{D}}$. Substituting $\nabla^2 \mathcal{D}_k(\mathbf{0}) = \overline{\boldsymbol{F}}_k$ into Eq. (55) proves Eq. (4). The cubic order follows from local Hessian Lipschitzness or, equivalently here, a locally bounded third derivative (Nesterov, 2013; Nocedal & Wright, 2006).

For two directions $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$, define $Q(s,t) := \mathcal{D}_k(s\boldsymbol{\delta}_1 + t\boldsymbol{\delta}_2)$. The chain rule gives

$$\left. \frac{\partial^2 Q}{\partial s \partial t} \right|_{s=t=0} = \boldsymbol{\delta}_1^\top \overline{\boldsymbol{F}}_k \boldsymbol{\delta}_2. \tag{57}$$

Equivalently, polarization of the Fisher quadratic form yields

$$2\boldsymbol{\delta}_1^\top \overline{\boldsymbol{F}}_k \boldsymbol{\delta}_2 = \|\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2\|_{\overline{\boldsymbol{F}}_k}^2 - \|\boldsymbol{\delta}_1\|_{\overline{\boldsymbol{F}}_k}^2 - \|\boldsymbol{\delta}_2\|_{\overline{\boldsymbol{F}}_k}^2, \tag{58}$$

where $\|\boldsymbol{\delta}\|_{\overline{\boldsymbol{F}}_k}^2 := \boldsymbol{\delta}^\top \overline{\boldsymbol{F}}_k \boldsymbol{\delta}$.

D CHECKPOINT OBSERVABILITY OF MIXED OUTPUT INTERACTION

This section derives the preceding and current base footprints on the same realized token coordinates and proves that their product recovers the mixed output interaction under the empirical output Fisher up to cubic local error.

D.1 TAYLOR EXPANSION OF THE PRECEDING CHECKPOINT FOOTPRINT

Recall $\boldsymbol{\theta}_{k-1} = \boldsymbol{\theta}_k - \boldsymbol{\delta}_{k-1}$. For each realized coordinate,

$$u_i = \ell_i(\boldsymbol{\theta}_k) - \ell_i(\boldsymbol{\theta}_k - \boldsymbol{\delta}_{k-1}). \tag{59}$$

Taylor expansion of $\ell_i(\boldsymbol{\theta}_k - \boldsymbol{\delta}_{k-1})$ around $\boldsymbol{\theta}_k$ gives

$$\ell_i(\boldsymbol{\theta}_k - \boldsymbol{\delta}_{k-1}) = \ell_i(\boldsymbol{\theta}_k) - \boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_{k-1} + \frac{1}{2}\boldsymbol{\delta}_{k-1}^\top \nabla^2 \ell_i(\boldsymbol{\theta}_k - t_i\boldsymbol{\delta}_{k-1})\boldsymbol{\delta}_{k-1} \tag{60}$$

for some $t_i \in (0,1)$. Hence

$$u_i = \boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_{k-1} + e_i^u, \tag{61}$$

where

$$e_i^u := -\frac{1}{2}\boldsymbol{\delta}_{k-1}^\top \nabla^2 \ell_i(\boldsymbol{\theta}_k - t_i\boldsymbol{\delta}_{k-1})\boldsymbol{\delta}_{k-1}, \qquad |e_i^u| \le \frac{L}{2}\|\boldsymbol{\delta}_{k-1}\|^2. \tag{62}$$

In vector form,

$$\boldsymbol{u} = \boldsymbol{\Phi}_k^\top \boldsymbol{\delta}_{k-1} + \boldsymbol{e}^u. \tag{63}$$

D.2 TAYLOR EXPANSION OF THE CURRENT BASE FOOTPRINT

The current base update ends at $\boldsymbol{\theta}_{k+1}^0 = \boldsymbol{\theta}_k + \boldsymbol{\delta}_k^0$. Therefore

$$d_i^0 = \ell_i(\boldsymbol{\theta}_k + \boldsymbol{\delta}_k^0) - \ell_i(\boldsymbol{\theta}_k). \tag{64}$$

Taylor expansion around $\boldsymbol{\theta}_k$ gives

$$d_i^0 = \boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_k^0 + e_i^d, \tag{65}$$

with

$$e_i^d := \frac{1}{2}(\boldsymbol{\delta}_k^0)^\top \nabla^2 \ell_i(\boldsymbol{\theta}_k + t_i'\boldsymbol{\delta}_k^0)\boldsymbol{\delta}_k^0, \qquad |e_i^d| \le \frac{L}{2}\|\boldsymbol{\delta}_k^0\|^2. \tag{66}$$

Thus

$$\boldsymbol{d}^0 = \boldsymbol{\Phi}_k^\top \boldsymbol{\delta}_k^0 + \boldsymbol{e}^d. \tag{67}$$

D.3 PROOF OF CHECKPOINT TOKEN–FISHER DUALITY

Multiplying Eqs. (61) and (65) coordinatewise gives

$$\begin{aligned} u_i d_i^0 = {} & (\boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_{k-1})(\boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_k^0) \\ & + (\boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_{k-1})e_i^d + (\boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_k^0)e_i^u + e_i^u e_i^d. \end{aligned} \tag{68}$$

Averaging the leading term,

$$\frac{1}{N}\sum_{i=1}^N (\boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_{k-1})(\boldsymbol{\varphi}_i^\top \boldsymbol{\delta}_k^0) = \boldsymbol{\delta}_{k-1}^\top \left(\frac{1}{N}\sum_{i=1}^N \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^\top\right) \boldsymbol{\delta}_k^0 = \boldsymbol{\delta}_{k-1}^\top \boldsymbol{F}_k \boldsymbol{\delta}_k^0. \tag{69}$$

Using $|\boldsymbol{\varphi}_i^\top \boldsymbol{\delta}| \le G\|\boldsymbol{\delta}\|$ and Eqs. (62)–(66), the total error satisfies

$$\begin{aligned} \left|\frac{1}{N}\boldsymbol{u}^\top \boldsymbol{d}^0 - \boldsymbol{\delta}_{k-1}^\top \boldsymbol{F}_k \boldsymbol{\delta}_k^0\right| \le {} & \frac{GL}{2}\|\boldsymbol{\delta}_{k-1}\|\|\boldsymbol{\delta}_k^0\|^2 \\ & + \frac{GL}{2}\|\boldsymbol{\delta}_k^0\|\|\boldsymbol{\delta}_{k-1}\|^2 \\ & + \frac{L^2}{4}\|\boldsymbol{\delta}_{k-1}\|^2\|\boldsymbol{\delta}_k^0\|^2. \end{aligned} \tag{70}$$

Under Eq. (44), the first two terms are $O(\eta^3)$ and the last is $O(\eta^4)$, proving Theorem 1. Equation (70) also states the finite-step approximation explicitly in addition to the asymptotic form.

D.4 Population and empirical Fisher

The checkpoint identity in Theorem 1 is deterministic in the realized empirical output metric $\boldsymbol{F}_k$ and requires no sampling assumptions. The KL Hessian is instead given by the population output Fisher $\overline{\boldsymbol{F}}_k$. We do not require $\boldsymbol{F}_k$ to be an unbiased estimator of $\overline{\boldsymbol{F}}_k$, particularly because the eligible coordinates are selected using the realized GRPO advantage. Thus, the empirical Fisher in our analysis should be interpreted as the realized metric appearing in the checkpoint identity rather than as a generic estimator of the objective Hessian (Kunstner et al., 2019).

D.5 Path cancellation and token-level backtracking

For arbitrary vectors $\boldsymbol{u}, \boldsymbol{d} \in \mathbb{R}^N$,

$$\|\boldsymbol{u}+\boldsymbol{d}\|^2 = \|\boldsymbol{u}\|^2 + \|\boldsymbol{d}\|^2 + 2\boldsymbol{u}^\top \boldsymbol{d}. \tag{71}$$

Rearranging proves Eq. (7). Its left-hand side is the sum of the two stepwise squared path lengths minus the squared net displacement, divided by two; it is positive when the two output motions exhibit net cancellation.

For the coordinatewise decomposition, define

$$\mathrm{Align}(\boldsymbol{u},\boldsymbol{d}) := \frac{1}{N}\sum_i [u_i d_i]_+, \qquad \mathrm{Rev}(\boldsymbol{u},\boldsymbol{d}) := \frac{1}{N}\sum_i [-u_i d_i]_+. \tag{72}$$

Since every scalar x satisfies $x = [x]_+ - [-x]_+$,

$$\frac{1}{N}\boldsymbol{u}^\top \boldsymbol{d} = \mathrm{Align}(\boldsymbol{u},\boldsymbol{d}) - \mathrm{Rev}(\boldsymbol{u},\boldsymbol{d}), \qquad \Gamma(\boldsymbol{u},\boldsymbol{d}) = \mathrm{Rev}(\boldsymbol{u},\boldsymbol{d}) - \mathrm{Align}(\boldsymbol{u},\boldsymbol{d}). \tag{73}$$

Therefore

$$[\Gamma(\boldsymbol{u},\boldsymbol{d})]_+ = [\mathrm{Rev} - \mathrm{Align}]_+ \le \mathrm{Rev}. \tag{74}$$

Token-level backtracking avoids cancellation between aligned and opposing coordinates. It decomposes into the two sign components

$$\mathrm{Rev}(\boldsymbol{u},\boldsymbol{d}) = \underbrace{\frac{1}{N}\sum_{i:u_i<0} (-u_i)[d_i]_+}_{\mathcal{B}^-(\boldsymbol{u},\boldsymbol{d})} + \underbrace{\frac{1}{N}\sum_{i:u_i>0} u_i[-d_i]_+}_{\mathcal{B}^+(\boldsymbol{u},\boldsymbol{d})}. \tag{75}$$

The first component is the negative-to-positive rebound controlled by OSOL.

E From Residual Coefficients to Backtracking Control

This section proves the control chain used in the main text: the constructed residual remains nonpositive after coefficient clipping, acts as a PPO likelihood-ratio penalty, induces a correction footprint through the token kernel, and contracts the targeted backtracking component.

E.1 Effective coefficient correction after clipping

Let $A_{\max} = 3$ denote the coefficient clipping bound in Eq. (21). The method constructs $r_i^{\mathrm{OSOL}} \le 0$ on $\mathcal{C}_k$ and $r_i^{\mathrm{OSOL}} = 0$ outside it; the complete rank and scale construction is given in Appendix F.3. Set

$$\begin{aligned} A_i^0 &:= \mathrm{clip}(A_i^{\mathrm{base}}, -A_{\max}, A_{\max}), \\ \widetilde{A}_i &:= \mathrm{clip}(A_i^{\mathrm{base}} + r_i^{\mathrm{OSOL}}, -A_{\max}, A_{\max}), \qquad \xi_i := \widetilde{A}_i - A_i^0. \end{aligned} \tag{76}$$

The scalar clipping map is monotone nondecreasing. Hence

$$\xi_i \le 0. \tag{77}$$

It is also 1-Lipschitz, so

$$|\xi_i| \le |r_i^{\mathrm{OSOL}}|. \tag{78}$$

Final coefficient clipping therefore never reverses the penalty direction and can only attenuate its magnitude. If $r_i^{\mathrm{OSOL}} = 0$, then $\xi_i = 0$.

E.2 Proof of the history-conditioned PPO penalty

Define the token log-ratio and likelihood ratio

$$z_i(\boldsymbol{\theta}) := \log \frac{\pi_{\boldsymbol{\theta}}(a_i \mid x_i)}{\pi_{\boldsymbol{\theta}_k}(a_i \mid x_i)}, \qquad \omega_i(\boldsymbol{\theta}) := e^{z_i(\boldsymbol{\theta})}. \tag{79}$$

Because rollouts are sampled at $\boldsymbol{\theta}_k$, the PPO reference policy is $\boldsymbol{\theta}_{\mathrm{old}} = \boldsymbol{\theta}_k$, so ω_i coincides with the ratio in Eq. (22). At the anchor, $z_i = 0$ and $\omega_i = 1$, which lies in the interior of the PPO clipping interval. In a local neighborhood, the token losses therefore reduce to the unclipped likelihood-ratio form (Schulman et al., 2017):

$$\mathcal{J}_i^0(z_i) = -e^{z_i} A_i^0, \qquad \mathcal{J}_i^{\mathrm{OSOL}}(z_i) = -e^{z_i}(A_i^0 + \xi_i). \tag{80}$$

Subtracting,

$$\mathcal{J}_i^{\mathrm{OSOL}}(z_i) - \mathcal{J}_i^0(z_i) = -e^{z_i}\xi_i = e^{z_i}|\xi_i| \ge 0. \tag{81}$$

Differentiating with respect to z_i and evaluating at the anchor gives

$$\left.\frac{\partial(\mathcal{J}_i^{\mathrm{OSOL}} - \mathcal{J}_i^0)}{\partial z_i}\right|_{z_i=0} = -\xi_i \ge 0, \tag{82}$$

which proves Lemma 1. Gradient descent therefore adds a penalty against increasing z_i on the selected support. The result is independent of the sign of A_i^0 because it concerns the incremental term induced by ξ_i.

E.3 Base update and correction footprint

Partition the response-token score matrix into focus and non-focus blocks, $\boldsymbol{\Phi}_{\mathrm{f}}$ and $\boldsymbol{\Phi}_{\mathrm{nf}}$. In the local unclipped policy-gradient model (Sutton et al., 1999; Schulman et al., 2017), the base mixed update is

$$\boldsymbol{\delta}_k^0 = \frac{\eta}{|\mathcal{R}|}\left(\boldsymbol{\Phi}_{\mathrm{f}}\boldsymbol{A}_{\mathrm{f}}^0 + \boldsymbol{\Phi}_{\mathrm{nf}}\boldsymbol{A}_{\mathrm{nf}}^0\right). \tag{83}$$

Let $\boldsymbol{\Phi}_{\mathcal{E}}$ contain the scores of eligible focus coordinates. Their first-order base footprint is

$$\begin{aligned} \boldsymbol{d}^0 &= \boldsymbol{\Phi}_{\mathcal{E}}^\top \boldsymbol{\delta}_k^0 + O(\eta^2) \\ &= \frac{\eta}{|\mathcal{R}|}\left(\boldsymbol{\Phi}_{\mathcal{E}}^\top \boldsymbol{\Phi}_{\mathrm{f}}\boldsymbol{A}_{\mathrm{f}}^0 + \boldsymbol{\Phi}_{\mathcal{E}}^\top \boldsymbol{\Phi}_{\mathrm{nf}}\boldsymbol{A}_{\mathrm{nf}}^0\right) + O(\eta^2). \end{aligned} \tag{84}$$

Thus $\boldsymbol{d}^0$ already includes the complete focus- and non-focus-domain contribution to the base update; the control argument does not remove or isolate non-focus contributions.

The effective OSOL correction is supported on eligible focus tokens:

$$\boldsymbol{\delta}_k^r = \frac{\eta}{|\mathcal{R}|}\boldsymbol{\Phi}_{\mathcal{E}}\boldsymbol{\xi}. \tag{85}$$

Its correction footprint is

$$\boldsymbol{h} := \boldsymbol{\Phi}_{\mathcal{E}}^\top \boldsymbol{\delta}_k^r = \frac{\eta}{|\mathcal{R}|}\boldsymbol{K}_{\mathcal{E}}\boldsymbol{\xi}, \qquad \boldsymbol{K}_{\mathcal{E}} := \boldsymbol{\Phi}_{\mathcal{E}}^\top \boldsymbol{\Phi}_{\mathcal{E}}. \tag{86}$$

This is the standard first-order function-space motion induced by a Jacobian Gram matrix (Jacot et al., 2018). Freezing the optimizer preconditioner at iteration k as a positive-semidefinite matrix P_k gives the local preconditioned response

$$\boldsymbol{h} = \frac{\eta}{|\mathcal{R}|}\boldsymbol{\Phi}_{\mathcal{E}}^\top P_k \boldsymbol{\Phi}_{\mathcal{E}}\boldsymbol{\xi}. \tag{87}$$

The preconditioned token kernel remains positive semidefinite. For AdamW, P_k approximates the local change in the optimizer update with its moment state frozen at iteration k; together with the unclipped PPO model at the anchor, this captures the first-order effect of the correction rather than the full clipped PPO–AdamW trajectory.

E.4 Residual-to-output sign transfer

From Eq. (86), for $i \in \mathcal{C}_k$,

$$h_i = \frac{\eta}{|\mathcal{R}|}\left[(\boldsymbol{K}_{\mathcal{E}})_{ii}\xi_i + \sum_{j\in\mathcal{C}_k,\, j\neq i}(\boldsymbol{K}_{\mathcal{E}})_{ij}\xi_j\right]. \tag{88}$$

A sufficient local self-response condition is

$$(\boldsymbol{K}_{\mathcal{E}})_{ii}|\xi_i| \geq \left|\sum_{j\in\mathcal{C}_k,\, j\neq i}(\boldsymbol{K}_{\mathcal{E}})_{ij}\xi_j\right|. \tag{89}$$

Since $(\boldsymbol{K}_{\mathcal{E}})_{ii} = \|\boldsymbol{\varphi}_i\|^2 \geq 0$ and $\xi_i \leq 0$, the diagonal term is nonpositive. Under Eq. (89),

$$\begin{aligned}(\boldsymbol{K}_{\mathcal{E}})_{ii}\xi_i + \sum_{j\neq i}(\boldsymbol{K}_{\mathcal{E}})_{ij}\xi_j &\leq -(\boldsymbol{K}_{\mathcal{E}})_{ii}|\xi_i| + \left|\sum_{j\neq i}(\boldsymbol{K}_{\mathcal{E}})_{ij}\xi_j\right| \\ &\leq 0,\end{aligned} \tag{90}$$

which proves Lemma 2. The condition holds automatically for a diagonal token kernel and under weighted diagonal dominance on the selected coordinates.

Coordinatewise nonpositivity is sufficient but not necessary. Using

$$[x+y]_+ - [x]_+ = \int_0^1 \mathbf{1}\{x+ty>0\}\, y\, dt \tag{91}$$

for almost every (x, y), the exact first-order backtracking change is

$$\begin{aligned}&\mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0 + \boldsymbol{h}) - \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0) \\ &\quad = \frac{1}{N}\int_0^1 \sum_{i\in\mathcal{C}_k}(-u_i)\mathbf{1}\{d_i^0 + th_i > 0\}h_i\, dt.\end{aligned} \tag{92}$$

Hence the weaker aggregate condition

$$\sum_{i\in\mathcal{C}_k}(-u_i)\mathbf{1}\{d_i^0 + th_i > 0\}h_i \leq 0, \qquad t\in[0,1], \tag{93}$$

is sufficient for contraction even if a small number of individual h_i are positive because of cross-token coupling.

E.5 Proof of first-order backtracking contraction

For $i \in \mathcal{C}_k$, $u_i < 0$. Under $h_i \le 0$, monotonicity of the positive-part function gives

$$[d_i^0 + h_i]_+ \le [d_i^0]_+. \tag{94}$$

Multiplying by $-u_i > 0$, summing, and dividing by N yields

$$\mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0 + \boldsymbol{h}) \le \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0). \tag{95}$$

The nonnegative difference is exactly the control gain in Eq. (12). For the selected-coordinate signed interaction,

$$\begin{aligned} &\frac{1}{N}\sum_{i\in\mathcal{C}_k} u_i(d_i^0 + h_i) - \frac{1}{N}\sum_{i\in\mathcal{C}_k} u_i d_i^0 \\ &\qquad = \frac{1}{N}\sum_{i\in\mathcal{C}_k} u_i h_i \ge 0, \end{aligned} \tag{96}$$

because both factors are nonpositive. Strict inequality holds whenever $u_i < 0$ and $h_i < 0$ on at least one selected coordinate.

E.6 Finite-step control gain

The local proof uses first-order footprints. The corrected footprint is

$$\boldsymbol{d}^{\mathrm{OSOL}} := \boldsymbol{\ell}_k(\boldsymbol{\theta}_k + \boldsymbol{\delta}_k^0 + \boldsymbol{\delta}_k^r) - \boldsymbol{\ell}_k(\boldsymbol{\theta}_k). \tag{97}$$

Taylor expansion around $\boldsymbol{\theta}_k$ gives

$$\boldsymbol{d}^{\mathrm{OSOL}} = \boldsymbol{d}^0 + \boldsymbol{h} + \boldsymbol{e}^r, \qquad \|\boldsymbol{e}^r\|_\infty = O(\eta^2). \tag{98}$$

Since $x \mapsto [x]_+$ is 1-Lipschitz,

$$\begin{aligned} &\left|\mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^{\mathrm{OSOL}}) - \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0 + \boldsymbol{h})\right| \\ &\quad \le \frac{1}{N}\sum_{i\in\mathcal{C}_k} (-u_i)|e_i^r| \le \frac{\|\boldsymbol{u}\|_1}{N}\|\boldsymbol{e}^r\|_\infty = O(\eta^3), \end{aligned} \tag{99}$$

because $\|\boldsymbol{u}\|_1/N = O(\eta)$. Therefore

$$\begin{aligned} \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^{\mathrm{OSOL}}) &\le \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0 + \boldsymbol{h}) + O(\eta^3) \\ &= \mathcal{B}^-(\boldsymbol{u}, \boldsymbol{d}^0) - \Delta_k^{\mathrm{ctrl}} + O(\eta^3), \end{aligned} \tag{100}$$

which proves Eq. (13). This form retains the controller's first-order reduction explicitly and separates it from the cubic deviation incurred by observing a finite checkpoint step.

F Design Principles and Complete Residual Construction

The preceding sections establish how a nonpositive effective coefficient becomes a backtracking-reducing output correction. This section derives the three design choices used to construct that coefficient: penalty-only support, rank-based allocation, and batchwise scale normalization.

F.1 Why OSOL uses the penalty-only support

The full tokenwise backtracking statistic is symmetric in the sign patterns $u_i < 0, d_i > 0$ and $u_i > 0, d_i < 0$. A symmetric controller would apply a nonpositive correction after negative drift and a nonnegative correction after positive drift. OSOL restricts the correction to the negative-drift candidate set because a nonpositive correction has a monotone backtracking-reduction property there and does not introduce a positive reward based only on checkpoint history.

For an arbitrary nonpositive output-space correction $v_i \le 0$, define

$$b_i(v_i) := [-u_i(d_i + v_i)]_+. \tag{101}$$

If $u_i < 0$,

$$b_i(v_i) = (-u_i)[d_i + v_i]_+ \le (-u_i)[d_i]_+ = b_i(0). \tag{102}$$

If $u_i > 0$,

$$b_i(v_i) = u_i[-d_i - v_i]_+ \ge u_i[-d_i]_+ = b_i(0). \tag{103}$$

Thus a nonpositive correction is backtracking-reducing on $\{u_i < 0\}$ but can increase backtracking on $\{u_i > 0\}$.

For any $\boldsymbol{v} \preceq 0$, define its negative-drift projection

$$v_i^- := \begin{cases} v_i, & u_i < 0, \\ 0, & u_i \ge 0. \end{cases} \tag{104}$$

Applying Eqs. (102)–(103) coordinatewise gives

$$\mathrm{Rev}(\boldsymbol{u}, \boldsymbol{d} + \boldsymbol{v}^-) \le \mathrm{Rev}(\boldsymbol{u}, \boldsymbol{d} + \boldsymbol{v}). \tag{105}$$

This identifies $\{u_i < 0\}$ as the support on which a nonpositive correction cannot increase backtracking; Lemma 2 connects this support to the implemented coefficient correction.

Suppressing the complementary pattern $u_i > 0, d_i < 0$ would require a positive coefficient correction $\xi_i^+ > 0$. At the PPO anchor,

$$\mathcal{J}_i^+ - \mathcal{J}_i^0 = -e^{z_i}\xi_i^+, \qquad \left.\frac{\partial(\mathcal{J}_i^+ - \mathcal{J}_i^0)}{\partial z_i}\right|_{z_i=0} = -\xi_i^+ < 0, \tag{106}$$

so gradient descent treats the historical signal as an additional reward-like incentive. A positive history correction may also oppose the current GRPO coefficient. We therefore use checkpoint history only as a one-sided penalty, preserving the current reward as the sole source of positive reinforcement.

F.2 Rank-optimal local allocation

Write a nonpositive output correction as $v_i = -e_i$ with $e_i \ge 0$. For $u_i < 0$, the exact reduction in scalar backtracking is

$$\begin{aligned} [-u_i d_i]_+ - [-u_i(d_i - e_i)]_+ &= (-u_i)\left([d_i]_+ - [d_i - e_i]_+\right) \\ &= (-u_i)\min\{e_i, [d_i]_+\}. \end{aligned} \tag{107}$$

When $d_i > 0$ and $0 \le e_i \le d_i$, the reduction is linear:

$$[-u_i d_i]_+ - [-u_i(d_i - e_i)]_+ = (-u_i)e_i, \tag{108}$$

so the marginal reduction per unit penalty is $-u_i$.

Proposition 2 (Rank-optimal local allocation)**.** *Consider active rebound coordinates with $u_i < 0$ and $d_i > 0$. Let $\{e_i\}$ be a fixed multiset of unsaturated penalty levels satisfying $0 \le e_i \le \min_j d_j$, so every level is feasible on every active coordinate. The total local reduction*

$$\Delta\mathcal{B}^- = \frac{1}{N}\sum_i (-u_i)e_i \tag{109}$$

is maximized by assigning larger e_i to more negative u_i.

Proof. Let $-u_i \geq -u_j$ but $e_i < e_j$. Swapping the two penalty levels changes the objective by

$$\frac{1}{N}\left[(-u_i)e_j + (-u_j)e_i - (-u_i)e_i - (-u_j)e_j\right] = \frac{1}{N}\big((-u_i) - (-u_j)\big)(e_j - e_i) \geq 0. \tag{110}$$

Repeatedly eliminating such inversions produces a co-monotone assignment without decreasing the reduction. Hence stronger penalties should be assigned to more negative preceding drift. □

Proposition 2 supplies the allocation principle behind Eq. (17). The quantile map implements the optimal ordering without using raw drift magnitude, is robust to isolated extremes, and separates relative allocation from the common scale chosen by τ. Under the diagonal self-response approximation $h_i \simeq \alpha_i \xi_i$ with positive, locally comparable α_i, the same ordering transfers from coefficient penalties to the correction footprint.

F.3 Complete residual construction

The base coefficient is

$$A_i^{\text{base}} = c_i A_i^{\text{g}}, \qquad c_i = \frac{w_{\text{src}(i)}}{Z_k}, \tag{111}$$

where Z_k is given by Eq. (15). The history residual is not multiplied by c_i again. The eligible set $\mathcal{E}_k$ contains response-token coordinates from the current focus domain whose trajectory-level GRPO advantage is nonzero, and

$$\mathcal{C}_k = \{i \in \mathcal{E}_k : u_i < 0\}. \tag{112}$$

For $n_k := |\mathcal{C}_k| > 0$, sort candidate drifts in ascending order and assign

$$p_i = \frac{\text{rank}_0(u_i) + 1/2}{n_k}. \tag{113}$$

The clipped affine profile is

$$\bar{r}_i = r_{\min} + \frac{\text{clip}(p_i, p_{\text{lo}}, p_{\text{hi}}) - p_{\text{lo}}}{p_{\text{hi}} - p_{\text{lo}}}(r_{\max} - r_{\min}), \tag{114}$$

with $\bar{r}_i = 0$ off $\mathcal{C}_k$. Since ascending rank is monotone,

$$u_i \leq u_j < 0 \quad \Longrightarrow \quad \bar{r}_i \leq \bar{r}_j \leq 0. \tag{115}$$

For $(p_{\text{lo}}, p_{\text{hi}}, r_{\min}, r_{\max}) = (0.1, 1, -1, 0)$,

$$\bar{r}_i = -1 + \frac{\text{clip}(p_i, 0.1, 1) - 0.1}{0.9}. \tag{116}$$

The implementation retains an explicit global gate $g_k \geq 0$ and focus mask $\lambda_i \in [0, 1]$:

$$q_i = g_k \lambda_i \bar{r}_i. \tag{117}$$

In the standard configuration, $g_k = 1$ and $\lambda_i = 1$ on the eligible focus coordinates, so Eq. (117) reduces to the conceptual profile in Eq. (18). The explicit terms are retained in code to support scheduling and masking without changing the core method.

Policy objective. After forming $\widetilde{A}_i$ in Eq. (21), define

$$\omega_i(\boldsymbol{\theta}) := \frac{\pi_{\boldsymbol{\theta}}(a_i \mid x_i)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_i \mid x_i)}. \tag{118}$$

The controller uses the unchanged clipped surrogate (Schulman et al., 2017):

$$\mathcal{L}_{\text{OSOL}}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{1}{|\mathcal{R}|}\sum_{i \in \mathcal{R}} \min\Big(\omega_i(\boldsymbol{\theta})\widetilde{A}_i, \text{clip}(\omega_i(\boldsymbol{\theta}), 1 - \epsilon_c, 1 + \epsilon_c)\widetilde{A}_i\Big)\right]. \tag{119}$$

No auxiliary objective is introduced; OSOL changes only the coefficient supplied to the existing policy update.

F.4 Adaptive scale identity and degenerate cases

Let

$$\kappa_k = \tau \frac{\sigma_A}{\sigma_q}, \qquad \sigma_A := \mathrm{std}_{\mathcal{E}_k}(A^{\mathrm{base}}), \qquad \sigma_q := \mathrm{std}_{\mathcal{E}_k}(q). \tag{120}$$

When σ_A or σ_q is nonfinite, or σ_q falls below the implementation threshold, the method sets $\kappa_k = 0$. Otherwise,

$$r_i^{\mathrm{OSOL}} = \kappa_k q_i = \kappa_k g_k \lambda_i \bar{r}_i \le 0, \tag{121}$$

and homogeneity of the standard deviation gives

$$\mathrm{std}_{\mathcal{E}_k}(r^{\mathrm{OSOL}}) = \kappa_k \mathrm{std}_{\mathcal{E}_k}(q) = \tau \mathrm{std}_{\mathcal{E}_k}(A^{\mathrm{base}}). \tag{122}$$

Thus

$$\frac{\mathrm{std}_{\mathcal{E}_k}(r^{\mathrm{OSOL}})}{\mathrm{std}_{\mathcal{E}_k}(A^{\mathrm{base}})} = \tau. \tag{123}$$

The identity is a batchwise scale normalization: τ fixes residual dispersion relative to base-advantage dispersion. Because noncandidate zeros remain in the eligible-set statistic, a sparse candidate set concentrates that dispersion on fewer coordinates, whereas a dense set distributes it more broadly. If $\mathcal{C}_k = \varnothing$ or the scale statistics are degenerate, $\kappa_k = 0$ and OSOL reduces exactly to the base mixed-domain objective.

F.5 Main-Training Reproducibility Details

Table 4 summarizes the training, generation, and OSOL-specific settings used in the main experiments. Let $\mathcal{D}$ denote the four training domains. At trainer step k, the focus domain is sampled uniformly from the domains other than the preceding focus,

$$f_k \sim \mathrm{Unif}(\mathcal{D} \setminus \{f_{k-1}\}). \tag{124}$$

The no-repeat constraint prevents one domain from leading two consecutive updates, while random selection avoids imposing a fixed domain order. All four domains remain in the mixed-domain batch. The sampled focus domain receives a base weight of 1.0, while each non-focus domain receives a weight of 0.3; only focus-domain response tokens are eligible for history correction. Each trainer step contains 128 prompts and eight sampled responses per prompt, yielding 1,024 rollout trajectories before optimization. Unless otherwise specified, the optimization and rollout settings follow the standard GRPO configuration in verl v0.5.0.

The preceding checkpoint is used only to rescore the same realized focus-domain token coordinates sampled by the current policy. All checkpoint-derived quantities are treated as fixed coefficients during optimization, and gradients are not propagated through θ_{k-1}. The resulting history residual is added to the task-weighted GRPO coefficient before applying the standard clipped policy objective.

Table 4: **Main-training and generation settings**. The prompt batch size is counted before rollout expansion.

Setting	Configuration
Model backbones	Qwen3-30B-A3B and Qwen3-8B-Base
Training domains	Mathematics, code generation, instruction following, and open-ended Chat
Training examples	7.5K examples per domain
Focus-domain schedule	Uniform random sampling from $\mathcal{D}\setminus\{f_{k-1}\}$; consecutive steps cannot use the same focus domain
Base domain weights	Focus domain: 1.0; each non-focus domain: 0.3
Prompt batch size	128 prompts per trainer step
Rollouts per prompt	8
Rollout trajectories per step	1,024
Training duration	15 epochs
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01
Learning rate	1×10^{-6} with a constant schedule and no warmup
PPO epochs	1 per rollout batch
PPO clip ratio	0.2 (symmetric clipping to $[0.8, 1.2]$)
Sampling temperature	1.0
Top-p / Top-k	1.0 / -1 (no nucleus or top-k truncation)
Maximum response length	8,192 tokens
Chat reward model	Skywork-Reward-Llama-3.1-8B-v0.2, following the Chat reward configuration used by MGS
History checkpoint	The frozen preceding checkpoint θ_{k-1}
Eligible history tokens	Focus-domain response tokens with nonzero trajectory-level GRPO advantage
Negative-drift candidates	Tokens satisfying $u_i < 0$
Rank-profile parameters	$(p_{\mathrm{lo}}, p_{\mathrm{hi}}, r_{\min}, r_{\max}) = (0.1, 1, -1, 0)$
Residual scale	$\tau = 0.03$
Final coefficient range	$[-3, 3]$