ON THE GEOMETRY OF ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) improves language-model reasoning, but the geometry of its attention weight updates remains poorly understood. We compare OPD with supervised fine-tuning (SFT) and reinforcement learning with verifiable rewards (RLVR). Endpoint diagnostics place OPD in a *relaxed off-principal regime*: more selective and spectrum-preserving than SFT, but less constrained than RLVR, with qualitative replication on Llama-3.2-3B. Our central trajectory-level finding is *subspace locking*: cumulative OPD updates align early with a persistent low-dimensional channel. Constraining subsequent training to an early rank-16 subspace leaves OPD performance essentially unchanged, whereas SFT degrades by a much larger margin. Token sparsification and off-policy rollouts preserve the rank trajectory, whereas changes to the objective signal affect its geometry. Mean-broadcast and objective-mixing controls further probe this sensitivity, while exposing confounding by training collapse and unequal gradient scales. Together, these findings characterize OPD's update trajectory and its functional role without positing a universal locking mechanism or claiming efficiency gains.

1 INTRODUCTION

Large reasoning models have substantially advanced mathematical and programming capabilities through post-training (Guo et al., 2025; Shao et al., 2024; OpenAI, 2024). Alongside supervised fine-tuning (SFT) on offline demonstrations (Wei et al., 2022) and reinforcement learning with verifiable rewards (RLVR) (Yu et al., 2025), on-policy distillation (OPD) trains a student on its own sampled trajectories with dense token-level feedback from a stronger teacher (Agarwal et al., 2024; Lu & Lab, 2025).

Despite its empirical utility, the geometry of OPD's weight updates remains poorly understood. SFT produces dense, principal-aligned updates, whereas RLVR makes sparse, off-principal changes that better preserve pretrained spectral structure (Liu et al., 2026; Mukherjee et al., 2025; Zhu et al., 2025). OPD combines dense teacher supervision with on-policy policy-gradient optimization, so neither paradigm alone predicts its update trajectory. We organize the analysis around three questions: where does OPD lie relative to SFT and RLVR, how does its update subspace evolve, and which components affect that evolution?

Locating OPD (Section 3). Using the endpoint diagnostics of Zhu et al. (2025), we find that OPD is more selective and geometry-preserving than SFT, but less constrained than RLVR: a *relaxed off-principal regime*. This ordering persists across Qwen scales, code data, and a Llama-3.2-3B replication that extends the endpoint evidence beyond one family. We use the Three-Gate account as an interpretation, not an OPD-specific proof.

Tracking the update channel (Section 4). Our central finding is *subspace locking*: OPD's cumulative updates rapidly enter a narrow, persistent low-dimensional channel. Update norms rule out a vanishing-update explanation, and subspace similarity shows early alignment with the final update directions. Constraining later training to an early rank-16 subspace leaves OPD essentially unchanged while degrading SFT, linking the geometry to functional sufficiency.

Testing the signal (Section 5). Token sparsification and off-policy rollouts preserve the OPD-like rank trajectory. Objective mixing alters its dynamics, although the quantified mixtures remain OPD-like at the shared checkpoint and do not establish a sharp transition. A mean-broadcast control shifts the endpoint diagnostics toward RLVR but also collapses in performance, limiting causal

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.

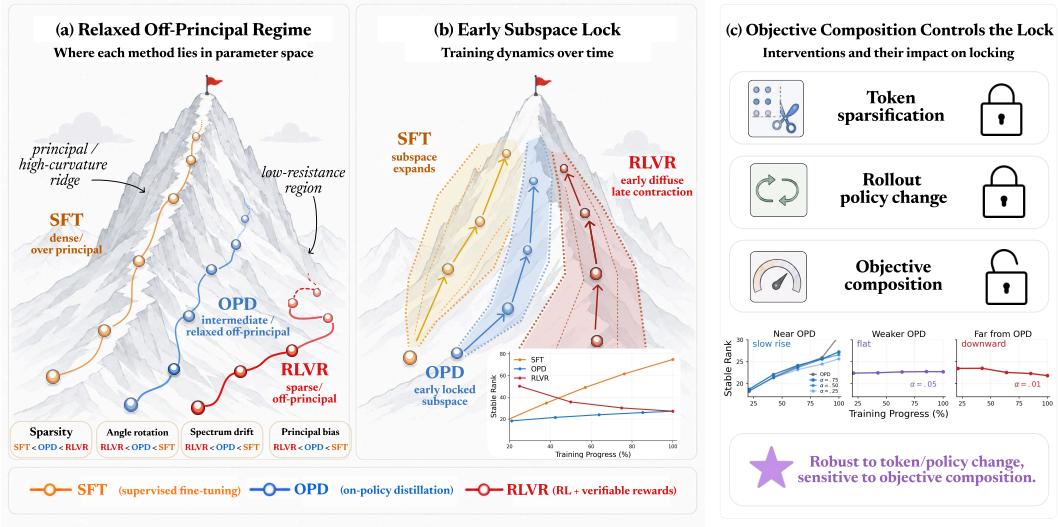

Figure 1: Overview of OPD update geometry. (a) OPD occupies a relaxed off-principal regime between SFT and RLVR. (b) Its cumulative updates align early with a low-dimensional subspace. (c) The observed trajectory is robust to token and rollout perturbations, but sensitive to changes in the objective signal.

interpretation. Together the controls distinguish supervision density from token-specific credit and motivate geometry-aware design.

2 Preliminaries

2.1 Notation

An analyzed weight matrix is $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. We write $\mathbf{W}_0$ for the initialization of a training stage, $\mathbf{W}_t$ for the weights after t optimizer steps, and

$$\Delta \mathbf{W}_t = \mathbf{W}_t - \mathbf{W}_0 \tag{1}$$

for the cumulative update. Matrices are set in bold uppercase ($\mathbf{W}, \mathbf{U}, \mathbf{\Sigma}, \mathbf{V}$), vectors in bold lowercase, and scalars in italics; $\mathbf{1}[\cdot]$ denotes the indicator. Checkpoints are indexed by optimizer step, and trajectories are compared at normalized training progress t/T; the checkpoint grids are listed in Appendix C. Unless stated otherwise, all quantities below are computed on the attention weight matrices of a single backbone—36 layers with four projection types each, 144 matrices per checkpoint—and our claims are confined to this set.

2.2 Post-training objectives

Supervised fine-tuning (SFT) maximizes the likelihood of offline demonstrations $(x, y) \sim \mathcal{D}$, so every token of every demonstration contributes to the update. Reinforcement learning with verifiable rewards (RLVR) samples rollouts from the current policy and optimizes an outcome reward; in the GRPO-style implementation we use, a group of G responses receives verifiable rewards $r_1, \ldots, r_G$ and the advantage

$$A_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)} \tag{2}$$

is broadcast over all tokens of response i (Shao et al., 2024). On-policy distillation (OPD) keeps the on-policy rollouts but replaces the scalar reward with dense teacher supervision: each generated token receives a coefficient $A_{i,t}$ derived from the teacher–student log-probability gap, so the response-level advantage of RLVR becomes token-dependent (Agarwal et al., 2024; Lu & Lab, 2025).

It is convenient to write the gradient contribution of one response in a common score-weighted form,

$$g = \sum_t a_t \, \phi_t, \qquad \phi_t = \nabla_\theta \log p_\theta(y_t \mid x, y_{<t}), \tag{3}$$

where only the coefficients a_t differ across objectives: constant within a response for SFT and RLVR, token-dependent for OPD. The controls in Section 5 perturb how these coefficients are produced.

2.3 SPECTRAL QUANTITIES

Let $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a singular value decomposition, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots$ and top-k truncations $\mathbf{U}_k, \mathbf{V}_k$. Principal angles compare the top-k left or right singular subspaces of two matrices,

$$\cos\theta_i^U = \sigma_i\big(\mathbf{U}_{0,k}^\top \mathbf{U}_{+,k}\big)\,, \tag{4}$$

and the normalized spectral shift (NSS) measures the relative movement of the spectrum,

$$\mathrm{NSS} = \frac{\|\sigma(\mathbf{W}_+) - \sigma(\mathbf{W}_0)\|_2}{\|\sigma(\mathbf{W}_0)\|_2}. \tag{5}$$

For a cumulative update we summarize effective dimensionality by the stable rank

$$\mathrm{srank}(\Delta\mathbf{W}_t) = \frac{\|\Delta\mathbf{W}_t\|_F^2}{\|\Delta\mathbf{W}_t\|_{\mathrm{op}}^2}, \tag{6}$$

and orientation by the alignment of the top-K right singular subspace with the final one,

$$\mathrm{Sim}_K(t, t_{\mathrm{end}}) = \frac{1}{K}\left\|\mathbf{V}_K(t)^\top \mathbf{V}_K(t_{\mathrm{end}})\right\|_F^2, \tag{7}$$

where values near one indicate shared singular directions. All quantities are computed per matrix and then averaged over the analyzed set; exact conventions are in Appendix E and Appendix C.3.

3 LOCATING OPD IN PARAMETER SPACE

Experimental setup. We report each run's configuration in full so that the comparison can be reproduced and its differences in compute are explicit. Our main comparison uses Qwen3-8B (Yang et al., 2025). OPD and RLVR share an SFT anchor and math-domain prompt distribution; the SFT update is measured from the pretrained base model to that anchor. Additional runs vary teacher/student scale, dense versus MoE teachers, and math versus code data. Training budgets differ across methods—SFT 5,375 steps at batch 256, OPD 299 at batch 64, and RLVR 1,023 at batch 512—so endpoints are selected by evaluation convergence instead of matched compute. Appendix C lists the full settings, save intervals, and the AIME plateau used to choose each horizon, and Section 4 provides trajectory-level evidence that does not rely on endpoint alignment.

3.1 PARAMETER-SPACE DIAGNOSTICS

We adopt four complementary diagnostics from Zhu et al. (2025). Let $\mathbf{W}_0$ and $\mathbf{W}_+$ denote the stage initialization and trained weights, with $\Delta\mathbf{W} = \mathbf{W}_+ - \mathbf{W}_0$; principal angles, NSS, stable rank, and subspace similarity are defined in Section 2. All geometric quantities below are computed on attention weight matrices—36 layers with four projection types each, 144 matrices per checkpoint; MLP, embedding, and output-head parameters are not analyzed, and our claims are confined to this set. The four diagnostics rule out different trivial accounts: differing update density, rotation of the pretrained subspaces, global spectral reshaping, and concentration on principal or high-magnitude weights.

Support and spectral structure. The bf16-aware sparsity is the fraction of unchanged entries after casting weights to bf16, using the relative tolerance $|w_+ - w_0| \leq 10^{-3}\max(|w_0|, |w_+|)$. Larger sparsity means fewer visible changes. Because both checkpoints are cast to bf16 before the comparison, the tolerance acts on values that are already bf16-quantized; we keep $\eta = 10^{-3}$ fixed as a convention and do not claim invariance to it. Smaller principal angles and NSS (Section 2) indicate stronger spectral preservation.

Update localization. For the bf16-visible update mask M, we report $|M \cap M_\star|/|M|$. The principal mask M_{princ} selects the top-α entries of a rank-k reconstruction of $\mathbf{W}_0$, while M_{low} selects the bottom-α entries by $|\mathbf{W}_0|$. Their random-overlap baseline is α. These masks characterize routing, not direct curvature measurements. Exact conventions and aggregation rules are in Appendix C.3.

Table 1: bf16-aware update sparsity. OPD lies between SFT and RLVR in the controlled Qwen3-8B comparison, and remains stable across OPD variants. Published checkpoints are used only as external reference points. In OPD variant names, TxB denotes a teacher model with xB parameters, e.g., OPD-4B-T8B uses a 4B student and an 8B teacher. Mixed† denotes a mixture of math, code, STEM, logic, and instruction data.

Base Model	**Finetuned (FT) Model**	**Algorithm**	**Data**	**sparsity**$_{\text{bf16}}$
Controlled comparison: SFT < OPD < RLVR				
Qwen3-8B-Base	Qwen3-8B-SFT	SFT	Math	8.1%
Qwen3-8B-SFT	OPD-8B-T32B	**OPD**	Math	**51.6%**
Qwen3-8B-SFT	RLVR-8B	GRPO	Math	77.2%
OPD robustness across teacher / student / data				
Qwen3-4B-SFT	OPD-4B-T8B	**OPD**	Math	**50.3%**
Qwen3-4B-SFT	OPD-4B-T14B	**OPD**	Math	**51.1%**
Qwen3-4B-SFT	OPD-4B-T32B	**OPD**	Math	**51.7%**
Qwen3-14B-SFT	OPD-14B-T32B	**OPD**	Math	**56.6%**
Qwen3-8B-SFT	OPD-8B-T32B-Code	**OPD**	Code	**57.1%**
Qwen3-8B-SFT	OPD-8B-T32B-MoE	**OPD**	Math	**48.6%**
Published reference points				
Qwen3-8B-Base	Klear-Reasoner-8B-SFT	SFT	Math+Code	0.6%
Qwen3-14B-Base	UniReason-14B-think-SFT	SFT	Mixed†	18.8%
Klear-Reasoner-8B-SFT	Klear-Reasoner-8B	GRPO	Math+Code	69.9%
Qwen3-8B-Base	GT-Qwen3-8B-Base	GRPO	Math	79.0%
Qwen3-4B	Polaris-4B-Preview	DAPO	Math	79.9%

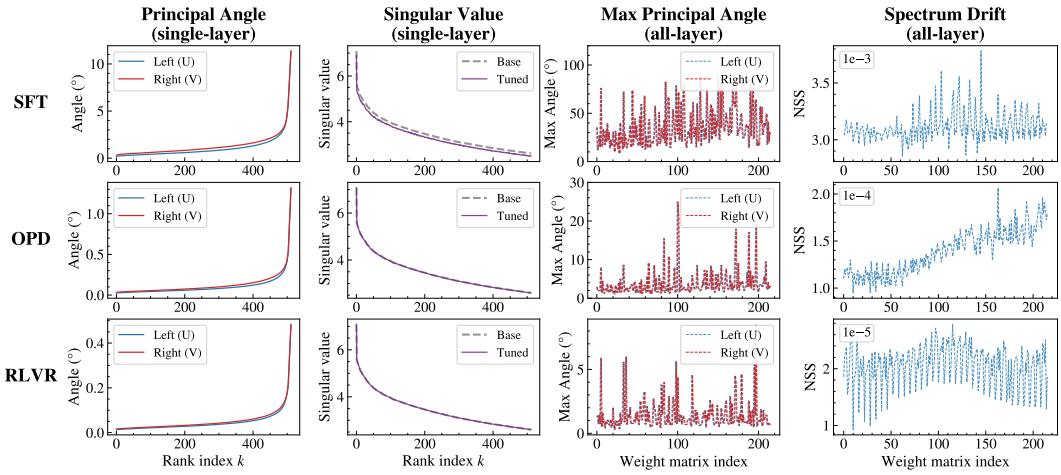

Figure 2: Parameter-space diagnostics for SFT, OPD, and RLVR. OPD lies between SFT's larger subspace rotations/spectral drift and RLVR's stronger preservation. Single-layer panels index singular directions by k; all-layer panels enumerate weight matrices across layers and modules.

3.2 A Relaxed Off-Principal Regime Across Settings

Qwen comparison. SFT leaves 8.1% of weights unchanged, OPD 51.6%, and RLVR 77.2%. Across teacher/student scales, code data, and an MoE teacher, OPD sparsity remains within 48.6–57.1% (Table 1). Spectral diagnostics agree: representative-layer angles exceed 10° for SFT, are around 1° for OPD, and below 0.5° for RLVR; NSS is respectively at the 10^{-3}, 10^{-4}, and 10^{-5} levels (Figure 2).

The localization analysis gives the same qualitative ordering. In the mask visualization, low-magnitude overlap increases monotonically across methods, 31.88%/53.59%/73.48% for SFT/OPD/RLVR, while principal-mask overlap is lowest for RLVR and highest for SFT (Figure 3); mask definitions, the retention fraction, and the corresponding random baseline are given

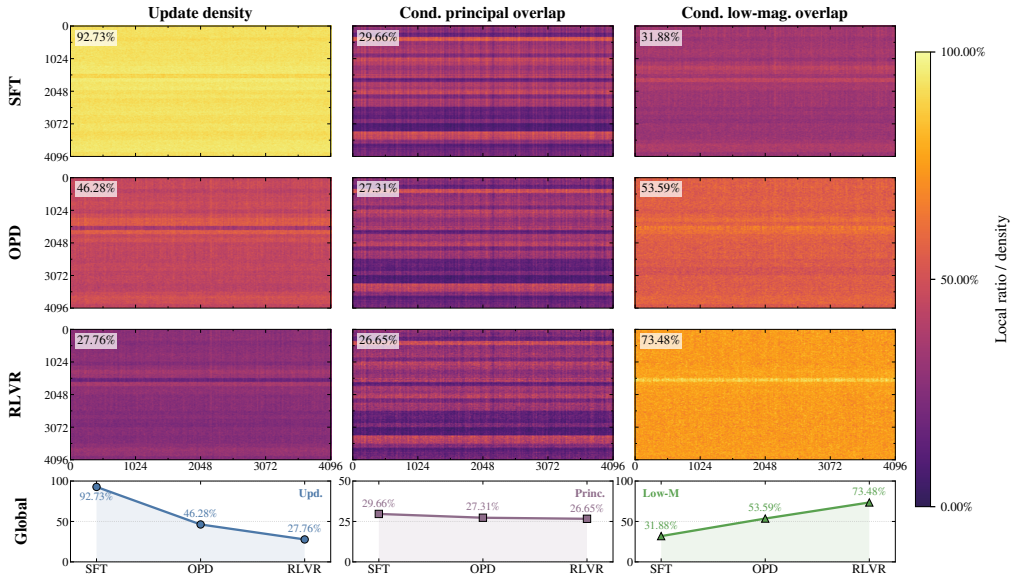

Figure 3: Update-mask localization. We compare where bf16-visible updates land relative to principal and low-magnitude masks. OPD shifts updates away from principal weights and toward low-magnitude regions, while remaining less selective than RLVR.

Table 2: Llama-3.2-3B endpoint geometry and the mean-broadcast control (Section 5.2). Rotation is the mean–median principal angle; routing is M_{low} at $\alpha = 0.1$. Broadcast values are rounded; its stable rank is not reported.

Method	Sparsity (%)	NSS	Rotation (°)	Routing	Stable rank
SFT	4.4	2.97×10^{-3}	1.74	0.105	98.2
OPD	46.4	1.48×10^{-4}	0.107	0.186	26.7
Mean-broadcast	≈ 70	2.3×10^{-5}	0.035	0.31	–
RLVR	84.6	4.76×10^{-6}	0.0115	0.565	40.4

in Appendix C.3. OPD therefore follows RLVR's off-principal bias with broader visible support: a *relaxed off-principal* regime.

Beyond Qwen. Llama-3.2-3B reproduces the intermediate OPD ordering on all four endpoint axes (Table 2). Its OPD update also has the lowest stable rank, consistent with a concentrated update but not, by itself, evidence of temporal locking. Backbone-specific training choices make this a qualitative replication instead of a compute-matched comparison. The main Qwen code-domain run likewise remains close to math-domain OPD, with sparsity 57.1% versus 51.6%, NSS 1.40×10^{-4} versus 1.39×10^{-4}, and stable rank 24.5 versus 27.1. Both domains remain reasoning-oriented.

3.3 An Organizing Three-Gate View

The Three-Gate account (Zhu et al., 2025) separates a distributional anchor, pretrained geometric routing, and bf16 realization. Writing the update as $g = \sum_t a_t \phi_t$ with $\phi_t = \nabla_\theta \log p_\theta(y_t \mid x, y_{<t})$, RLVR broadcasts a sequence-level coefficient whereas OPD uses token-dependent teacher–student discrepancies; this motivates testing whether token-specific credit changes the update geometry. The covariance decompositions in Appendix D organize those interventions and state the measurable quantity they identify.

4 Subspace Locking in OPD

Endpoint concentration does not establish a stable training channel. We therefore track cumulative updates $\Delta\mathbf{W}_t = \mathbf{W}_t - \mathbf{W}_0$ across checkpoints and separately test their dimension, orientation, and functional role.

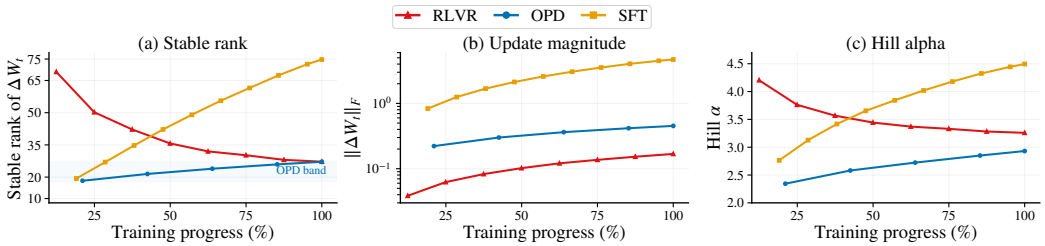

Figure 4: Intrinsic update geometry. (a) OPD stays in a narrow stable-rank band while SFT expands and RLVR contracts. (b) OPD accumulates a larger update than RLVR despite comparable endpoint stable rank. (c) Hill estimates provide an auxiliary spectral-shape diagnostic.

4.1 A Low-Dimensional Update Band

For each matrix we use the stable rank of the cumulative update (Section 2), averaged across analyzed matrices. Figure 4(a) shows that OPD stays in a narrow low-dimensional band, whereas SFT expands and RLVR contracts toward its endpoint. This is a trajectory-level distinction, not an interpolation of endpoint ranks. We use *locking* to mean two concrete conditions over the observed window: the stable rank of the cumulative update stays inside a band rather than trending, and the top-16 subspace similarity to the final update satisfies $\mathrm{Sim}_{16}(t, t_{\mathrm{end}}) \geq \tau$ after an early t_0, with $\tau \approx 0.7$ in our runs; OPD meets both from its first measured checkpoint (21% of training). Checkpoint grids differ across methods (Appendix C), so comparisons use normalized progress.

Low rank is not explained by negligible parameter movement: on Llama, OPD ends at a lower stable rank than RLVR (26.7 vs. 40.4, Table 2) yet accumulates a larger Frobenius norm (Figure 4(b)). To characterize spectral shape we use the Hill tail estimate of the top singular values,

$$\widehat{\alpha}_{\mathrm{Hill}} = \Big[\frac{1}{k-1}\sum_{i=1}^{k-1}\log\frac{\sigma_i}{\sigma_k}\Big]^{-1}, \tag{8}$$

which increases as the spectrum becomes heavier-tailed. It evolves mildly for OPD, increases for SFT, and decreases for RLVR (panel (c)). Because the three statistics are computed on the same matrices and averaged in the same way, they jointly describe a concentrated spectrum of a nontrivial update rather than a degenerate one; exact conventions are in Appendix E.

4.2 Stability and Functional Sufficiency

Early subspace emergence. A low-rank update could still rotate through changing subspaces. Let $\mathbf{V}_K(t)$ contain the top-K right singular vectors of $\Delta\mathbf{W}_t$. We measure alignment with the final update channel by subspace similarity (Section 2). At $K = 16$, OPD is already aligned at the first measured checkpoint; SFT and RLVR approach their final subspaces more gradually (Figure 5). Thus, OPD's low-dimensional channel emerges early and persists.

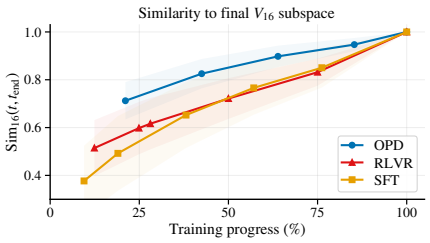

Figure 5: Early subspace emergence. Top-16 subspace similarity to the final update rises earlier for OPD than for SFT or RLVR.

Functional sufficiency. For both OPD and SFT, we extract $\mathbf{V}_{16}$ from the cumulative update around 20% training progress and constrain subsequent gradient matrices to that early right subspace, $g \leftarrow g\mathbf{V}_{16}\mathbf{V}_{16}^{\top}$. Rank 16 is a stringent bottleneck, close to but slightly below the early effective dimension; we do not claim it is optimal. Figure 6 shows a large asymmetry: SFT degrades substantially over the matched window, whereas the OPD difference is a few points on AIME 2024—a 30-problem benchmark whose evaluation noise is of comparable order—so we read the result as a bottleneck-sensitivity difference instead of a precise retention estimate. The intervention establishes functional sufficiency in the tested setting, not computational savings from full-gradient projection.

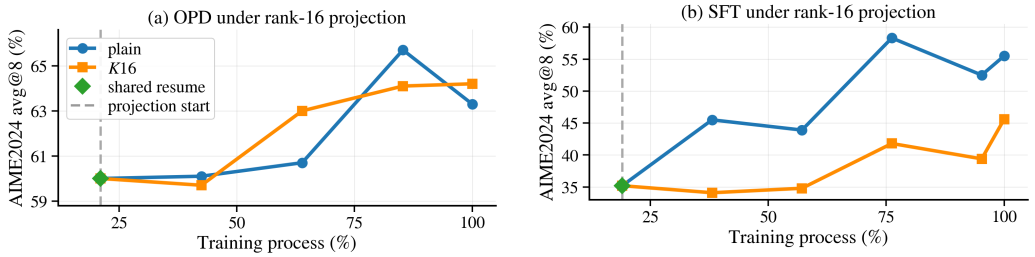

Figure 6: Rank-16 projected continuation for (a) OPD and (b) SFT. Projection into a subspace extracted around 20% training progress preserves OPD performance but degrades SFT.

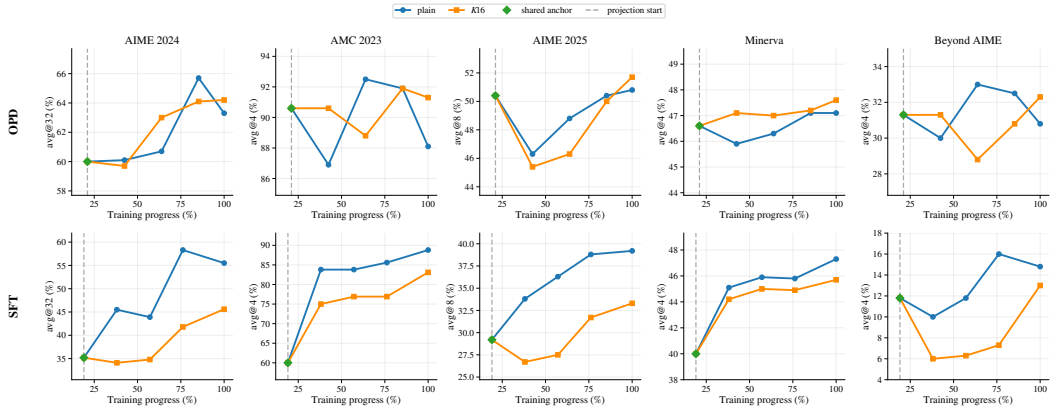

Figure 7: Additional evaluation of rank-constrained training. Unconstrained and rank-16 projected training across five reasoning benchmarks, using the same early subspace, projection start, and matched windows as Figure 6. Green diamonds mark the shared anchor checkpoint and dashed vertical lines mark the projection start.

Beyond the primary evaluation. To check that this pattern is not specific to AIME 2024, we repeat both projected runs on five reasoning benchmarks with the same subspace, projection start, and matched windows. Figure 7 shows that OPD is consistently less affected by the bottleneck than SFT. Appendix F records the protocol; as before, the comparison identifies neither an optimal rank nor any wall-clock benefit from projecting full gradients.

5 What Controls Subspace Locking?

We perturb supervision density, rollout policy, and the objective signal to probe the channel identified in Section 4. Stable rank is the primary trajectory diagnostic; Appendix G provides protocols and auxiliary metrics.

5.1 Runtime Perturbations Preserve the Lock

Retaining either top-KL or randomly selected tokens at 25% or 50% density leaves the OPD stable-rank trajectory nearly unchanged (Figure 8(a)). Even random 25% retention changes update scale more than spectral shape, so the low-dimensional profile is not confined to a small set of high-KL tokens. Teacher-generated, off-policy rollouts likewise preserve stable rank despite a modestly larger update norm (panel (b)). These interventions alter coverage or sampled trajectories while retaining token-dependent teacher feedback.

5.2 Token-Specific Credit: A Mean-Broadcast Control

Token thinning still retains token-specific coefficients. To test their role more directly, we write $A_{i,t} = \bar{A}_i + \widetilde{A}_{i,t}$, with $\bar{A}_i$ the mean over valid response tokens. On Llama-3.2-3B, we broadcast

Table 3: Controlled variables in Section 5. Each intervention changes one target factor while keeping the remaining OPD setup fixed where possible.

Dimension	Baseline	Token sparse	Off-policy	Alpha mix
Rollout policy	Student	Student	Teacher	Student
Token coverage	100%	25/50%	100%	100%
Objective signal	OPD	OPD	OPD	OPD/RLVR mix
Reward module	OPD	OPD	OPD	Mixed reward
Other settings	–	Same	Same	Same

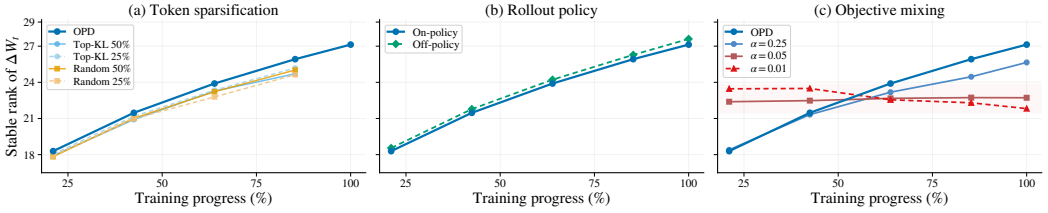

Figure 8: Stable-rank trajectories under (a) token sparsification, (b) rollout-policy changes, and (c) objective mixing. Runtime perturbations preserve the OPD-like profile; mixing changes its dynamics. The displayed trajectories do not establish a critical mixture coefficient.

$\bar{A}_i$ to every valid token and otherwise retain the OPD setup. This removes within-response credit variation while preserving the response-level mean.

All four endpoint diagnostics move from OPD toward RLVR (Table 2): sparsity rises to about 70%, NSS falls to 2.3×10^{-5}, rotation to $0.035°$, and low-magnitude routing rises to 0.31. However, AIME accuracy also approaches zero and the truncation ratio reaches about 0.9. Mean broadcasting therefore changes geometry and training behavior jointly, so it is not a clean isolation of credit variation; together with token thinning it distinguishes retaining a token-dependent signal on fewer positions from removing its within-response variation.

5.3 Objective Mixing and Its Quantitative Limits

We mix the native OPD and RLVR advantages in the GRPO-style implementation:

$$A_i^{(\alpha)} = \alpha A_{i,\mathrm{OPD}} + (1-\alpha) A_{i,\mathrm{RLVR}}. \tag{9}$$

Unlike thinning or changing rollouts, this alters the gradient source (Figure 8(c)). To quantify proximity to the pure-objective trajectories at matched progress, let D_{OPD} and D_{RLVR} be the mean absolute stable-rank distances. We use

$$S_{\mathrm{OPD}}(\alpha) = \frac{D_{\mathrm{RLVR}}(\alpha)}{D_{\mathrm{OPD}}(\alpha) + D_{\mathrm{RLVR}}(\alpha)}, \tag{10}$$

where larger values mean closer to OPD. At the common step-63 checkpoint, the four quantified mixtures are strongly OPD-like (Table 4). The intervals resample 144 weight matrices within a run, not independent seeds. These results locate neither a sharp transition nor a critical α; the table does not quantify the $\alpha = 0.01$ curve of Figure 8.

Direction and scale are coupled. After normalized-progress alignment, pure OPD and RLVR have median gradient norms 0.456 and 0.139, a ratio of about $3.3\times$, with stage-wise ratios of $3.8\times$, $2.9\times$, and $3.2\times$ for early, middle, and late normalized progress. Nominal α is therefore not the effective fractional contribution of each objective. Our results describe native mixing; norm-matched mixing would be a different, dynamically reweighted algorithm. The covariance account in Appendix G.3 motivates this sensitivity but does not predict a transition, and denser sweeps and independent seeds remain necessary.

What the controls separate. Read together, the perturbations constrain explanations instead of settling them. Token thinning changes which tokens carry supervision, and off-policy rollouts

Table 4: OPD-likeness at the common step-63 checkpoint. Intervals are 95% matrix-bootstrap intervals over 144 weight matrices, not independent training seeds.

α	0.75	0.50	0.25	0.05
S_{OPD}	0.94	0.94	0.94	0.99
95% interval	[0.93,0.96]	[0.93,0.95]	[0.93,0.95]	[0.96,1.00]

change how they were sampled; neither moves the rank trajectory, whereas removing within-response credit variation or mixing the objective does. Each intervention carries its own confound—mean broadcasting collapses training, and mixing is confounded by unequal gradient scales—so we report the objective signal as implicated, not identified.

6 Discussion & Conclusion

OPD occupies an intermediate off-principal endpoint regime while following an early, persistent low-dimensional update channel. The rank-constrained continuation test links this trajectory to functional sufficiency, going beyond endpoint diagnostics inherited from PNT (Zhu et al., 2025). Cross-backbone evidence broadens the empirical scope, and the controls distinguish token coverage, token-specific credit, and native objective composition.

Where this points. The most direct algorithmic direction is reduced-space continuation after a full-parameter warm-up: because an early rank-16 channel is sufficient for OPD but not for SFT, continuing training inside a fixed rank-K subspace is plausible for OPD and risky for methods whose updates expand into new directions. Stable rank and subspace similarity are inexpensive to monitor, and among the factors we perturb it is the objective composition that moves them. Comparisons across mixtures should be norm-matched, since a nominal coefficient is not the effective contribution when gradient scales differ by roughly $3.3\times$. These are testable directions, not validated recipes: projecting full gradients demonstrates neither memory nor wall-clock savings, and neither an optimal rank nor broad generalization is established.

What would change the picture. Our controls vary token coverage, rollout policy, and the objective while holding the anchor and data fixed. Holding the objective fixed and varying the anchor would separate anchor- and data-driven accounts from objective-driven ones, and a stronger test would predict the channel's emergence rather than measure it afterwards. We report locking as an empirical regularity of the settings studied here.

Limitations

The evidence covers Qwen-family reasoning runs and a qualitative Llama replication, not arbitrary models or tasks. All geometry is computed on attention weight matrices (144 per checkpoint); MLP, embedding, and output-head parameters are untested. Endpoints use unequal budgets, saved checkpoints only resolve dynamics at their sampling times, and we report single training runs per setting—the mixing intervals resample matrices within a run, not independent seeds. The bf16 tolerance is fixed at $\eta = 10^{-3}$ and its sensitivity is untested. Mean broadcasting is confounded by performance collapse. The Three-Gate and covariance accounts are hypothesis-generating, and locking denotes the persistence of an early channel in these settings rather than a guarantee for other objectives or anchors. Broader tasks, norm-matched mixtures, denser sweeps, random-subspace and orthogonal-complement controls, and efficient reduced-space implementations require separate validation.

AI Use Statement

We used ChatGPT for writing assistance, including language polishing, LaTeX formatting, organization suggestions, and refinement of presentation. We also used it to assist with drafting plotting and analysis scripts. All scientific claims, experimental designs, code, figures, numerical results, and final manuscript content were reviewed, verified, and edited by the authors. No generative AI system was used as an author, to produce experimental results, or to make autonomous scientific decisions.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Yuchen Cai, Ding Cao, Liang Lin, Chunxi Luo, Xin Xu, Kai Yang, Weijie Liu, Saiyong Yang, Tianxiang Zhao, Guangzhong Sun, Guiquan Liu, and Junfeng Fang. Learning to foresee: Unveiling the unlocking efficiency of on-policy distillation, 2026. URL https://arxiv.org/abs/2605.11739.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020. URL https://arxiv.org/abs/2002.06305.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes, 2026. URL https://arxiv.org/abs/2603.25562.

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

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2958370.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018. URL https://arxiv.org/abs/1801.06146.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026. URL https://arxiv.org/abs/2604.13016.

Zihang Liu, Tianyu Pang, Oleg Balabanov, Chaoqun Yang, Tianjin Huang, Lu Yin, Yaoqing Yang, and Shiwei Liu. Lift the veil for the truth: Principal weights emerge after rank reduction for reasoning-focused supervised fine-tuning, 2026. URL https://arxiv.org/abs/2506.00772.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level, 2025. URL https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-mini-Level-1cf81902c14680b3bee5eb349a512a51. Notion Blog.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models, 2025. URL https://arxiv.org/abs/2505.11711.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael

Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

OpenAI. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. Rl's razor: Why online reinforcement learning forgets less, 2025. URL https://arxiv.org/abs/2509.04259.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models, 2026. URL https://arxiv.org/abs/2604.00626.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. URL https://arxiv.org/abs/2109.01652.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin, 2026. URL https://arxiv.org/abs/2507.14843.

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

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

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, David Z. Pan, Zhangyang Wang, Yuandong Tian, and Kai Sheng Tai. The path not taken: RLVR provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A Artifact Use

We use publicly available model checkpoints, datasets, and benchmarks only for research evaluation and analysis. We cite the corresponding creators in the main text or appendix and use these artifacts consistently with their intended research use. We checked the public license or usage terms of the major artifacts where available, including Qwen3, DAPO-Math-17k, Dolci-Think SFT, DeepCoder, and LiveCodeBench. We do not redistribute the original artifacts.

B Reproducibility Statement

We specify the training setup, data, and endpoint-selection protocol for every reported run in Appendix C, and the OPD variant configurations in Appendix C.2; published checkpoints used only as external reference points are labeled as such in Table 1. The parameter-space diagnostics are defined in Appendix C.3 and the trajectory metrics in Appendix E. Control experiments share one documented baseline, with perturbation protocols and controlled variables listed in Appendix G. All artifacts we build on are publicly available, and their use is described in Appendix A.

C Experimental Details

C.1 Training Setup

This section summarizes the training setups used in the main parameter-space analyses. Table 5 gives the shared setup. Tables 6, 7, and 8 report the default supervised finetuning (SFT), on-policy distillation (OPD), and reinforcement learning with verifiable rewards (RLVR) settings, respectively. OPD and RLVR are initialized from the same SFT anchor, while the SFT trajectory is measured from the pretrained base checkpoint. We use Qwen3 checkpoints (Yang et al., 2025); math-domain OPD/RLVR runs use dapo-math-17k (Yu et al., 2025), and the SFT anchor is trained on Dolci-Think SFT data (Olmo et al., 2025).

Table 5: Shared settings used in the main training runs.

Item	Setting	Item	Setting
Model family	Qwen3	Main backbone	Qwen3-8B
Evaluation	AIME 2024 avg@8	Precision	bf16 training
Gradient accumulation	fp32	Attention softmax	fp32
Hardware	H200 80GB GPUs	Dropout	0.0

Table 6: Default SFT setting used to produce the shared anchor checkpoint.

Item	Setting	Item	Setting
Initialization	Qwen3-8B-Base	Objective	Supervised next-token prediction
Loss mask	Qwen3 loss mask	Training data	Dolci-Think SFT data
Epochs	5	Final checkpoint	`iter_0005375`
Global batch size	256	Learning rate	10^{-5}
Schedule	Cosine decay to 10^{-6}	Warmup	10% of training steps
Optimizer	Adam	Weight decay	0.1
Adam betas	$(0.9, 0.95)$	Save interval	256 steps
Eval interval	256 steps	Role	Produces the shared SFT anchor

Endpoint selection. We choose the main analyzed endpoints according to the evaluation curves observed during training. The SFT, OPD, and RLVR runs showed little additional improvement near the selected checkpoints, corresponding approximately to 5k, 300, and 1k training steps, respectively. We use these rounded horizons as the main comparison points for interpretability and reproducibility, instead of tuning endpoints to optimize the parameter-space diagnostics.

The evaluation-convergence rationale is performance-based rather than compute matching. OPD reaches an AIME 2024 plateau around 66% between steps 191 and 256, RLVR around 68% between

Table 7: Default OPD setting used in the main experiments.

Item	Setting	Item	Setting
Initialization	SFT anchor `iter_0005375`	Student	Qwen3-8B
Teacher	Qwen3-32B	Training data	dapo-math-17k
Rollout policy	Student-generated, on-policy	Rollout prompts	300 per step
Samples per prompt	8	Rollout temperature	1.0
Training steps	299	Save interval	64 steps
Objective	On-policy distillation	Teacher supervision	Token log-probabilities
OPD coefficient	1.0	Advantage estimator	GRPO
KL loss	Disabled	Entropy bonus	Disabled
Global batch size	64	Learning rate	10^{-6}
Schedule	Constant	Optimizer	Adam
Weight decay	0.1	Adam betas	$(0.9, 0.98)$

Table 8: Default RLVR setting used in the main experiments.

Item	Setting	Item	Setting
Initialization	SFT anchor `iter_0005375`	Student	Qwen3-8B
Teacher	None	Training data	dapo-math-17k
Rollout policy	Student-generated, on-policy	Rollout prompts	1024 per step
Samples per prompt	8	Rollout temperature	1.0
Training steps	1023	Objective	Verifier-reward GRPO
Reward	Binary DeepScaler math accuracy	Advantage estimator	GRPO
KL loss	Disabled	Entropy bonus	Disabled
PPO clipping	$\epsilon = 0.2$, $\epsilon_{\text{high}} = 0.28$	Dynamic filter	Skip batches with zero reward variance
Global batch size	512	Learning rate	10^{-6}
Schedule	Constant	Optimizer	Adam
Weight decay	0.1	Adam betas	$(0.9, 0.98)$
Adam epsilon	10^{-15}		

steps 768 and 896, and SFT around 60% by step 4,608. These observations motivate the respective horizons; unequal budgets remain a limitation, and the geometric trajectories provide evidence beyond a single endpoint.

Geometry origins. For OPD and RLVR, $\mathbf{W}_0$ is the shared SFT anchor `iter_0005375`. For SFT, $\mathbf{W}_0$ is the pretrained Qwen3-8B-Base checkpoint. All updates are computed as $\Delta\mathbf{W} = \mathbf{W}_+ - \mathbf{W}_0$ relative to the corresponding stage initialization. Thus, OPD and RLVR share a common origin in the parameter-space comparison.

C.2 OPD Variant Setup

All OPD variant runs follow Table 7 unless the changed factor is listed explicitly. Table 9 summarizes the student, teacher, and changed factor for each variant. Additional variant-specific notes are provided in Table 10.

The code-domain variant uses DeepCoder data (Luo et al., 2025); evaluation is conducted on LiveCodeBench v5 (Jain et al., 2024).

Table 9: OPD variant runs used for robustness checks in Table 1.

Variant	Student	Teacher	Changed factor
Baseline	Qwen3-8B	Qwen3-32B	Reference
MoE teacher	Qwen3-8B	Qwen3-30B-A3B	Teacher architecture
4B→8B	Qwen3-4B	Qwen3-8B	Teacher scale
4B→14B	Qwen3-4B	Qwen3-14B	Teacher scale
4B→32B	Qwen3-4B	Qwen3-32B	Teacher scale
14B→32B	Qwen3-14B	Qwen3-32B	Student scale
Code domain	Qwen3-8B	Qwen3-32B	Data domain
Multi-seed	Qwen3-8B	Qwen3-32B	Random seed

Table 10: Additional notes for OPD variant runs.

Item	Setting	Item	Setting
Code domain	Uses DeepCoder data and LiveCodeBench v5 evaluation	MoE teacher	Uses Qwen3-30B-A3B as the teacher model
Multi-seed	Uses the same Qwen3-8B/Qwen3-32B math setting with different random seeds		

C.3 Parameter-Space Diagnostic Implementation

All diagnostics are computed offline on saved checkpoints. Each analysis loads a pair of checkpoints $(\mathbf{W}_0, \mathbf{W}_+)$ and computes $\Delta\mathbf{W} = \mathbf{W}_+ - \mathbf{W}_0$. For OPD and RLVR, $\mathbf{W}_0$ is the shared SFT anchor; for SFT, $\mathbf{W}_0$ is the pretrained Qwen3-8B-Base model. Metrics are computed per matrix and then averaged across matrices unless otherwise specified.

bf16-aware update sparsity. Both $\mathbf{W}_0$ and $\mathbf{W}_+$ are cast to bf16 and then back to fp32 before comparison. A scalar entry is considered unchanged when

$$|\mathbf{W}^{\mathrm{bf16}}_{+,ij} - \mathbf{W}^{\mathrm{bf16}}_{0,ij}| \le \eta \max\big(|\mathbf{W}^{\mathrm{bf16}}_{0,ij}|, |\mathbf{W}^{\mathrm{bf16}}_{+,ij}|\big)\,,$$
$$\eta = 10^{-3}.$$

The bf16-aware sparsity is

$$\mathrm{sp}_{\mathrm{bf16}}(\mathbf{W}_0, \mathbf{W}_+) = 1 - \frac{1}{n}\sum_{i,j} \mathbf{1}[\mathbf{W}_{+,ij} \not\approx_\eta \mathbf{W}_{0,ij}]\,,$$

where n is the number of entries in the analyzed matrix. Overall sparsity is parameter-count weighted across analyzed weights.

We keep η fixed at 10^{-3} for all reported numbers and do not sweep it; the tolerance acts on values that are already bf16-quantized, so it should be read together with the realization threshold $\frac{1}{2}\mathrm{ULP}_{\mathrm{bf16}}$ above instead of as a separately tuned parameter.

Principal-angle rotation. For each matrix, let

$$\mathbf{W}_0 = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^\top, \mathbf{W}_+ = \mathbf{U}_+\mathbf{\Sigma}_+\mathbf{V}_+^\top.$$

Let $\mathbf{U}_{0,k}, \mathbf{V}_{0,k}$ and $\mathbf{U}_{+,k}, \mathbf{V}_{+,k}$ denote the top-k left and right singular subspaces. We compute principal angles by

$$\cos\theta_i^U = \sigma_i(\mathbf{U}_{0,k}^\top\mathbf{U}_{+,k}), \cos\theta_i^V = \sigma_i(\mathbf{V}_{0,k}^\top\mathbf{V}_{+,k}).$$

We use $k = 512$ by default and report summary statistics of the resulting angle vectors. These diagnostics use CPU float64 SVD, since GPU SVD can produce unstable angles for near-identical matrices.

Spectral drift. Let $\sigma(W)$ denote the singular-value vector of $\mathbf{W}$ in descending order. We measure normalized spectral shift as

$$\mathrm{NSS}(\mathbf{W}_0, \mathbf{W}_+) = \frac{\|\sigma(\mathbf{W}_+) - \sigma(\mathbf{W}_0)\|_2}{\|\sigma(\mathbf{W}_0)\|_2}.$$

NSS is computed per matrix and then averaged without parameter-count weighting.

Update–mask overlap. The visible update mask is

$$M_{\mathrm{upd}} = \{(i,j) : \mathbf{W}_{+,ij} \not\approx_\eta \mathbf{W}_{0,ij}\}.$$

We compare it with two masks derived from $\mathbf{W}_0$. The principal mask M_{princ} contains the top-α fraction of entries by magnitude in the rank-r SVD reconstruction of $\mathbf{W}_0$, and the low-magnitude mask M_{low} contains the bottom-α fraction of entries by $|\mathbf{W}_0|$. For $M_\star \in \{M_{\mathrm{princ}}, M_{\mathrm{low}}\}$, we report

$$\mathrm{Overlap}(M_\star, M_{\mathrm{upd}}) = \frac{|M_\star \cap M_{\mathrm{upd}}|}{|M_{\mathrm{upd}}|}.$$

We use rank $r = 64$ and $\alpha = 0.5$ unless otherwise specified. The random baseline is α, since each reference mask contains an α fraction of entries.

Matrix coverage and averaging. For bf16-aware sparsity, we compute overall sparsity over all analyzed weight parameters and report both overall and per-type summaries. For spectral geometry and update-mask overlap, we analyze standard weight matrices across layers; QKV projections are split into Q, K, and V when needed. Metrics are first computed per matrix and then averaged across matrices. This avoids domination by a small number of large matrices while preserving layer-wise and module-wise variation.

D Objectives and Three-Gate Extension

This section collects the organizing account and general mathematical tools used to motivate the experiments. The coefficient decompositions and perturbation bounds do not prove OPD-specific subspace locking. Statements about anchoring and geometric routing are interpretations under the stated local assumptions, not necessary or sufficient conditions for the observed trajectory.

Original Three-Gate account. The Three-Gate account of RLVR (Zhu et al., 2025) explains visible update sparsity as the outcome of a constrained, geometry-steered, and precision-filtered optimization process.

Gate I: KL anchor. RLVR updates are locally constrained in policy space. In a KL-regularized or trust-region view, the post-update policy remains close to the current or reference policy:

$$D_{\mathrm{KL}}(\pi_{\theta^+} \,\|\, \pi_\theta) \le K. \tag{11}$$

This policy-space leash prevents each step from freely rewriting the pretrained model.

Gate II: model geometry. The KL anchor limits the size of the update, but not its location. The second gate attributes this location to pretrained model geometry. Given a bounded step, the structured loss landscape steers updates away from high-curvature principal directions and toward lower-curvature, spectrum-preserving regions. This explains why RLVR updates are off-principal rather than merely small.

Gate III: precision realization. The final gate determines which coordinates become visible in stored weights. Under bf16 precision, sub-threshold micro-updates are not realized as changed parameters. Thus the underlying routing bias appears as high bf16-aware update sparsity: preferred regions receive visible updates, while changes elsewhere remain hidden.

Training objectives. We write the three post-training paradigms in a common notation. For an input x and sequence $y = (y_1, \ldots, y_T)$, SFT minimizes cross-entropy on offline demonstrations:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t=1}^{T} \log p_\theta(y_t^\star \mid x, y_{<t}^\star). \tag{12}$$

RLVR uses on-policy samples with a scalar sequence-level advantage:

$$\mathcal{L}_{\mathrm{RLVR}} = -A(y) \sum_{t=1}^{T} \log p_\theta(y_t \mid x, y_{<t}) + \beta \mathcal{R}_{\mathrm{KL}}, \tag{13}$$

where

$$\mathcal{R}_{\mathrm{KL}} = D_{\mathrm{KL}}(p_\theta(\cdot \mid x) \,\|\, p_{\mathrm{ref}}(\cdot \mid x)). \tag{14}$$

Here $A(y)$ denotes a normalized sequence-level advantage and p_{ref} is the reference policy. Clip-only RL variants can be viewed as replacing the explicit KL penalty with a local trust-region effect.

OPD optimizes student-generated rollouts against a teacher distribution:

$$\mathcal{L}_{\mathrm{OPD}} = \sum_{t=1}^{T} D_{\mathrm{KL}}\left(p_\theta^t \,\|\, q_T^t\right), \tag{15}$$

where

$$p_\theta^t = p_\theta(\cdot \mid x, y_{<t}), q_T^t = q_T(\cdot \mid x, y_{<t}).$$

If a forward-KL implementation is used, the KL direction is replaced accordingly; the signal-granularity argument below is unchanged.

Unified score-weighted form. Let

$$\phi_t = \nabla_\theta \log p_\theta(y_t \mid x, y_{<t}) \tag{16}$$

denote the token score. A broad class of stochastic post-training updates can be abstracted as

$$g = \sum_{t=1}^{T} a_t \phi_t. \tag{17}$$

The distinction lies in the coefficient a_t. In RLVR, $a_t = A(y)$ is shared across the sequence. In OPD, a_t varies across tokens, induced by teacher–student disagreement. Thus OPD keeps the on-policy structure of RLVR but replaces a scalar credit signal with dense token-level supervision.

OPD-specific relaxation. The relaxation appears at the level of update covariance. Let $C = \mathbb{E}[gg^\top]$. For RLVR,

$$C_{\mathrm{RLVR}} = \mathbb{E}\left[A(y)^2 s s^\top\right], s = \sum_t \phi_t. \tag{18}$$

For OPD,

$$C_{\mathrm{OPD}} = \mathbb{E}\left[\sum_{t,t'} a_t a_{t'} \phi_t \phi_{t'}^\top\right]. \tag{19}$$

RLVR couples all token scores through one sequence-level scalar. OPD uses heterogeneous token coefficients, which can change directional support within a geometry-steered update family. The decomposition alone does not guarantee that the support broadens. This is the sense in which OPD is *relaxed*: it remains anchored and geometry-constrained, but less selective than RLVR.

Local anchor and weight bound. We now make the anchoring step explicit. For a weight block $\mathbf{W}$, consider a local quadratic constraint

$$\frac{1}{2}\langle \mathrm{vec}(\Delta\mathbf{W}), S_{\mathbf{W}}\mathrm{vec}(\Delta\mathbf{W})\rangle \le \delta_W, S_{\mathbf{W}} \succeq \mu_{\mathbf{W}} I. \tag{20}$$

This implies

$$\|\Delta\mathbf{W}\|_F \le \sqrt{\frac{2\delta_{\mathbf{W}}}{\mu_{\mathbf{W}}}}. \tag{21}$$

The bound captures the anchoring effect. OPD differs from RLVR not by removing the anchor, but by using denser token-level information inside the anchored region.

Spectral stability under bounded updates. Let $\mathbf{W}_+ = \mathbf{W}_0 + \Delta\mathbf{W}$ and $\gamma_k = \sigma_k(\mathbf{W}_0) - \sigma_{k+1}(\mathbf{W}_0)$ be the singular-value gap. By Wedin-type perturbation bounds,

$$\begin{aligned}&\max\{\|\sin\Theta(\mathbf{U}_k(\mathbf{W}_0), \mathbf{U}_k(\mathbf{W}_+))\|_2,\\ &\|\sin\Theta(\mathbf{V}_k(\mathbf{W}_0), \mathbf{V}_k(\mathbf{W}_+))\|_2\} \le \frac{\|\Delta\mathbf{W}\|_2}{\gamma_k}.\end{aligned} \tag{22}$$

Singular values satisfy

$$|\sigma_i(\mathbf{W}_+) - \sigma_i(\mathbf{W}_0)| \le \|\Delta\mathbf{W}\|_2, \tag{23}$$

and

$$\|\sigma(\mathbf{W}_+) - \sigma(\mathbf{W}_0)\|_2 \le \|\Delta\mathbf{W}\|_F. \tag{24}$$

Therefore, updates with small operator and Frobenius norm tend to preserve the pretrained spectral structure. OPD remains in this bounded regime, but its dense token-level signal permits larger subspace rotation and spectral drift than RLVR.

Precision realization. bf16 precision determines which coordinates become visible in stored weights. A coordinate is realized as changed only if

$$|\Delta\mathbf{W}_{ij}| \gtrsim \tfrac{1}{2}\mathrm{ULP}_{\mathrm{bf16}}(\mathbf{W}_{0,ij}). \tag{25}$$

The corresponding visible update mask is

$$M_{ij} = \mathbf{1}\left[|\Delta\mathbf{W}_{ij}| \gtrsim \tfrac{1}{2}\mathrm{ULP}_{\mathrm{bf16}}(\mathbf{W}_{0,ij})\right]. \tag{26}$$

Since OPD updates are larger than RLVR updates but remain geometry-constrained, more coordinates cross the realization threshold, yielding intermediate bf16-aware sparsity.

What the account makes measurable. The decomposition identifies coefficient heterogeneity as the quantity that should govern the effective rank of the update: if the token coefficients a_t vary more within a response, the score-weighted sum can excite more directions, whereas broadcasting a single coefficient removes that variation. Our present controls do not isolate this quantity cleanly. Mean broadcasting also changes training stability, α-mixing changes the coefficient distribution and its scale at the same time, and RLVR differs from OPD in more than coefficient uniformity, so we state the trend as the quantity a dedicated experiment would need to vary while holding coverage, scale, and optimization fixed.

E Trajectory Metric Definitions

For trajectory-level analysis, we study the cumulative update $\Delta\mathbf{W}_t = \mathbf{W}_t - \mathbf{W}_0$ at each checkpoint. Unless otherwise specified, metrics are computed for each analyzed weight matrix and then averaged across matrices.

Stable rank. Stable rank measures the effective number of dominant singular directions that carry the update energy:

$$\mathrm{srank}(\Delta\mathbf{W}_t) = \frac{\|\Delta\mathbf{W}_t\|_F^2}{\|\Delta\mathbf{W}_t\|_{\mathrm{op}}^2}. \tag{27}$$

Unlike algebraic rank, stable rank is insensitive to arbitrarily small singular values and therefore provides a scale-aware measure of effective update dimensionality.

Frobenius norm. We use the Frobenius norm to measure cumulative update magnitude:

$$\|\Delta \mathbf{W}_t\|_F = \left(\sum_{i,j} (\Delta \mathbf{W}_{t,ij})^2 \right)^{1/2}. \tag{28}$$

This diagnostic rules out the possibility that low stable rank is merely caused by negligible parameter movement.

Hill tail estimate. To inspect the spectral shape of $\Delta \mathbf{W}_t$, we use the Hill tail estimator (Hill, 1975). Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$ denote the selected top singular values. We estimate the tail exponent as

$$\widehat{\alpha}_{\text{Hill}} = \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{\sigma_i}{\sigma_k} \right]^{-1}. \tag{29}$$

We use this only as an auxiliary spectral-shape diagnostic, complementary to stable rank.

F Additional Evaluation of Rank-Constrained Training

Figure 7 in the main text reports the full five-benchmark comparison. Both projected runs reuse the subspace extracted from the cumulative update around 20% of training, the same projection start, and the matched training windows of the primary evaluation; the only change relative to the unconstrained runs is the rank-16 constraint applied to subsequent gradient matrices.

G Control Experiment Details

G.1 Shared OPD Control Setup

The Qwen token-sparsification, off-policy, and objective-mixing experiments in Section 5 are based on the same OPD baseline configuration. The mean-broadcast experiment instead uses Llama-3.2-3B and is described separately below. Unless otherwise specified, each intervention changes only the stated factor and keeps the remaining OPD training pipeline fixed. Full training hyperparameters are provided in Appendix C.

Table 11: Shared setup for OPD control experiments.

Item	Setting	Item	Setting
Student model	Qwen3-8B	Teacher model	Qwen3-32B
Student initialization	Qwen3-8B SFT anchor, iter_0005375	Training data	dapo-math-17k
Training length	300 steps	Checkpoint steps	63, 127, 191, 255, 299
Baseline rollout	Student-generated, 8 samples per prompt	Analyzed matrices	36 layers $\times$ 4 attention weights
Primary diagnostic	Stable rank	Auxiliary diagnostics	Frobenius norm, Hill tail estimate

For all spectral diagnostics, we analyze the same matrix set as in the main trajectory analysis: 144 matrices, corresponding to 36 layers and four attention weight types per layer. Metrics are computed per matrix and then averaged across matrices.

G.2 Perturbation Protocols

Token sparsification. This intervention perturbs the density of token-level teacher supervision while keeping the rollout policy and OPD objective fixed. For each sampled response token a_t, the implementation computes the sampled-token student–teacher log-probability gap

$$r_t = \log \pi_\theta(a_t \mid s_t) - \log \pi_T(a_t \mid s_t),$$

where π_θ is the student policy and π_T is the teacher policy. This is a sampled-token log-probability gap, not a full-vocabulary KL divergence.

For token sparsification, we introduce a binary token mask $m_t \in \{0, 1\}$. In top-KL retention, tokens are ranked by r_t and the largest ρ fraction is retained; this selects tokens where the student is most over-confident relative to the teacher. In random retention, $m_t = 1$ for a uniformly sampled ρ fraction of response tokens. We use $\rho \in \{0.25, 0.50\}$.

The retained-token signal is compensated by

$$\widetilde{r}_t = \frac{T}{\sum_{\tau=1}^{T} m_\tau} m_t r_t.$$

The additive OPD update then uses this gap as a correction term, subtracting it from the advantage. Thus, the top-KL variant selects the tokens with the largest student–teacher deviation, while the training signal penalizes this deviation. Masking is applied independently for each sample at the response-token level. All other settings, including teacher model, on-policy rollout generation, optimizer, batch size, learning rate, and training data, are kept identical to the OPD baseline. The sparsification runs are evaluated at checkpoints $s = \{63, 127, 191, 255\}$.

Off-policy rollouts. This intervention perturbs the rollout policy while keeping the token-level OPD objective fixed. In the baseline, rollouts are generated by the current student policy:

$$y \sim \pi_\theta(\cdot \mid x).$$

In the off-policy variant, rollouts are generated by the teacher:

$$y \sim \pi_T(\cdot \mid x),$$

and the student then computes log-probabilities on these teacher-generated tokens during the training forward pass. The training objective remains pure OPD KL distillation; no reward mixing is introduced. Thus, this intervention changes the sampling distribution but keeps the teacher-token gradient source fixed.

The off-policy runs use Qwen3-32B teacher rollouts served through an sglang router with two teacher replicas. The importance-ratio clipping threshold is set to $\epsilon = 1.0$, making clipping effectively inactive for this comparison. The off-policy checkpoints are evaluated at $s = \{63, 127, 191, 255, 299\}$.

Mean-broadcast OPD on Llama. For each response i with valid token set V_i, we replace every valid token advantage with $\bar{A}_i = |V_i|^{-1} \sum_{t \in V_i} A_{i,t}$. The other settings of this Llama OPD run are unchanged. This experiment is separate from the Qwen hyperparameters above: the cross-backbone comparison required backbone-specific training choices and is used as qualitative evidence. Table 2 reports the available endpoint diagnostics; the broadcast run's AIME accuracy approaches zero and its truncation ratio reaches approximately 0.9. The concurrent collapse prevents geometry-only causal attribution.

Objective interpolation. This intervention perturbs the objective composition while keeping the student rollout policy fixed. We implement this at the advantage level by mixing the OPD teacher-correction signal with the RLVR reward signal:

$$A_i^{(\alpha)} = \alpha A_{i,\mathrm{OPD}} + (1 - \alpha) A_{i,\mathrm{RLVR}}, \tag{30}$$

where $A_{i,\mathrm{OPD}} = -r_i$ and $r_i = \log \pi_\theta(a_i \mid s_i) - \log \pi_T(a_i \mid s_i)$ is the sampled-token student–teacher log-probability gap. Thus, the OPD term encourages correction toward the teacher. $A_{i,\mathrm{RLVR}}$ is the GRPO advantage from the math accuracy reward, broadcast to response tokens. The resulting policy-gradient direction is

$$g_\alpha = \mathbb{E}_i \left[A_i^{(\alpha)} \nabla_\theta \log \pi_\theta(a_i \mid s_i) \right]. \tag{31}$$

We evaluate $\alpha \in \{0.75, 0.50, 0.25, 0.05, 0.01\}$, with $\alpha = 1$ corresponding to the pure OPD baseline and $\alpha = 0$ corresponding to the RLVR endpoint. All alpha-mixing runs use raw signal mixing without per-batch standard deviation normalization. The rollout policy remains on-policy, and the optimizer, batch size, learning rate, training data, and evaluation setup are identical to the baseline.

Mixing quantification and gradient scales. Table 4 reports the common step-63 comparison. The 95% intervals resample 144 analyzed matrices within the runs; they are not confidence intervals over independently trained seeds. After aligning the pure-OPD and pure-RLVR gradient-norm traces by normalized progress, overall medians are 0.456 and 0.139, respectively. The early-, middle-, and late-stage OPD/RLVR ratios are $3.8\times$, $2.9\times$, and $3.2\times$. The nominal mixture coefficient therefore weights native signals of unequal scale. Norm matching, denser coefficient sweeps, and independent seeds remain separate experiments, not results claimed here.

G.3 Conditional Covariance Interpretation

The following account is a conditional interpretation of the controls, not an OPD-specific theorem or a prediction of a sharp mixture boundary. The controls separate sample perturbations from objective perturbations. OPD gradients can be written as token-level sums,

$$g_{\mathrm{OPD}} = \sum_t J_t^\top \delta_t,\ \tilde{g}_{\mathrm{OPD}} = \sum_t m_t J_t^\top \delta_t, \tag{32}$$

where J_t is the local Jacobian, δ_t is the teacher–student token discrepancy, and m_t is a token mask. If token gradients share a dominant update subspace, masking primarily rescales the second moment,

$$\mathbb{E}[\tilde{g}\tilde{g}^\top] \approx c\,\mathbb{E}[g_{\mathrm{OPD}} g_{\mathrm{OPD}}^\top] + \mathrm{noise}, \tag{33}$$

so the leading spectral directions are preserved. Off-policy rollouts change the sampling distribution, but retain the same teacher-token gradient source. Objective mixing changes the source itself:

$$\begin{aligned} g_\alpha &= \alpha g_{\mathrm{OPD}} + (1-\alpha) g_{\mathrm{RLVR}}, \\ \mathbf{\Sigma}_\alpha &\approx \alpha^2 \mathbf{\Sigma}_{\mathrm{OPD}} + (1-\alpha)^2 \mathbf{\Sigma}_{\mathrm{RLVR}} \\ &\quad + \alpha(1-\alpha)\mathbf{\Sigma}_{\mathrm{cross}}. \end{aligned} \tag{34}$$

Changing the OPD coefficient can alter the dominant covariance geometry. This general decomposition is consistent with objective sensitivity, but does not predict the direction, magnitude, or sharpness of a transition; those require empirical measurements.

G.4 Auxiliary Metrics

Stable rank is the primary diagnostic in Section 5 because the question is whether the locked low-dimensional update channel is preserved. We additionally track update scale and spectral shape through Frobenius norm and Hill tail estimates (Figure 9). These auxiliary diagnostics verify that runtime perturbations preserve the OPD-like spectral trajectory up to scale shifts, whereas objective interpolation changes the update geometry more substantially.

H Related Work

Post-training and distillation. SFT learns from offline demonstrations (Howard & Ruder, 2018; Dodge et al., 2020; Wei et al., 2022); RLHF and RLVR optimize preference or outcome-level feedback (Ziegler et al., 2020; Ouyang et al., 2022; Shao et al., 2024; Guo et al., 2025). OPD instead provides dense teacher feedback on student-generated rollouts (Agarwal et al., 2024; Lu & Lab, 2025; Song & Zheng, 2026). Recent studies examine failure modes, scaling, and efficiency (Fu et al., 2026; Li et al., 2026; Cai et al., 2026). We complement these behavioral accounts with a parameter-space study of OPD's training trajectory.

Weight-space geometry. Prior work contrasts SFT's dense principal-direction updates (Liu et al., 2026) with RL's localized, off-principal changes (Mukherjee et al., 2025), and relates conservative RL updates to limited policy drift (Shenfeld et al., 2025; Wu et al., 2026). Most directly, PNT (Zhu et al., 2025) supplies the endpoint diagnostics and Three-Gate account adopted here: an update-size anchor, model-conditioned geometric routing, and finite-precision realization. We explicitly build on these tools, including their diagnostic and visualization conventions; our contribution is the OPD trajectory analysis, early subspace-locking observation, rank-constrained functional test, and objective-sensitive controls. The inherited diagnostic suite itself is not a new methodological contribution.

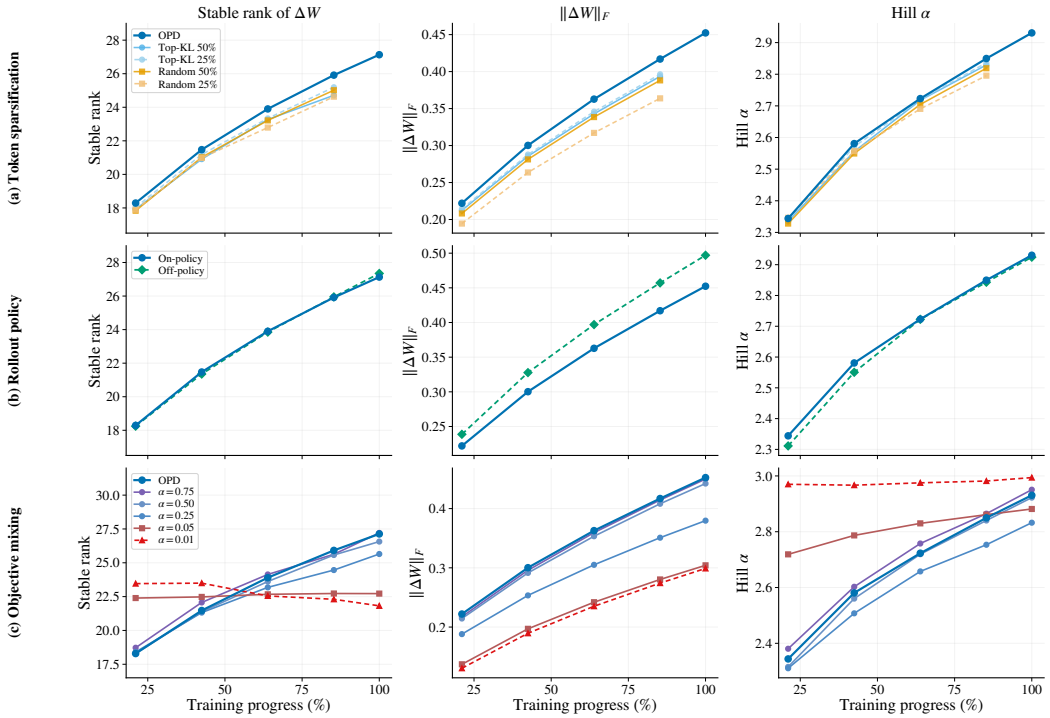

Figure 9: **Auxiliary metrics for control experiments.** We report update scale and spectral-shape diagnostics for the same perturbations analyzed in Figure 8. Runtime perturbations preserve the OPD-like spectral profile up to modest scale changes, whereas objective interpolation induces a distinct trajectory.