ON THE ORTHOGONAL GEOMETRY OF ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student on its own sampled trajectories using dense token-level teacher feedback. Standard low-rank adaptation (LoRA) makes OPD more parameter-efficient, but restricts learning to an initialization-dependent channel whose suitability for OPD is unclear. We introduce OP-OPD and OM-OPD, two orthonormal spectral initialization methods for low-rank OPD. They initialize the LoRA row space with, respectively, the principal and minor right-singular directions of pretrained weights, while a zero output factor preserves the initial model function. A first-step analysis shows that this choice projects the OPD gradient onto a selected subspace, making the initialization channel explicit. On three mathematical reasoning benchmarks, both methods improve over standard LoRA-OPD on aggregate mean@16 and maj@16 in two teacher–student settings. Update-trajectory and spectral diagnostics show that the two variants follow distinct pretrained channels, consistent with the proposed routing interpretation. The results identify initialization geometry as a design dimension for OPD, with gains observed across both evaluated teacher–student pairs.

1 INTRODUCTION

On-policy distillation (OPD) trains a student on trajectories sampled from its own policy while using a teacher to provide token-level guidance on those student-visited states (Hinton et al., 2015; Agarwal et al., 2024; Lu & Lab, 2025; Li et al., 2026). This combination makes the supervision distribution responsive to the student while retaining dense feedback at each generated token. A full-parameter OPD reference updates all model parameters, whereas LoRA represents each weight update with a trainable low-rank factorization while keeping the pretrained weight fixed (Hu et al., 2022).

The limitation of low-rank OPD is not necessarily rank alone. At a fixed rank, initialization selects the subspace through which learning first enters the adapter, yet standard LoRA leaves that channel unspecified (Lin et al., 2026; Xiao et al., 2026; Guo et al., 2025a). This raises a more specific question: can a deliberate initialization subspace help low-rank OPD use dense teacher guidance more effectively?

We introduce OP-OPD and OM-OPD to make the initialization channel an explicit design choice. They use equal-budget channels drawn from opposite ends of the pretrained spectrum while preserving the initial model function. Section 3 gives the projection analysis and the precise construction. Prior analyses identify early subspace locking in full-parameter OPD (Shen et al., 2026), but do not determine how to initialize a low-rank channel for OPD.

Across AIME25, AMC23, and AIME24, we compare the two spectral channels with standard LoRA-OPD in two teacher–student settings and analyze their training trajectories and update geometry (Guo et al., 2025b; Shao et al., 2024; Yu et al., 2025).

Our contributions are:

- We introduce OP-OPD and OM-OPD, two orthonormal spectral initialization methods that adapt low-rank updates to the token-level, on-policy supervision regime of OPD while preserving the initial model function.

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.

- We make the initialization-dependent gradient channel explicit. Under a first-step SGD analysis, row-orthonormality yields a projection of the OPD gradient, and the principal or minor pretrained subspace determines the projected component.
- We evaluate the two spectral channels against matched-rank standard LoRA-OPD on three reasoning benchmarks in two teacher–student settings, finding consistent aggregate improvements.

2 Related Work

On-policy distillation and update geometry. OPD uses student-generated trajectories and teacher-provided token-level supervision (Agarwal et al., 2024; Lu & Lab, 2025; Niu et al., 2025). This line of work primarily changes the supervision seen by the student: which self-generated states are distilled, how reasoning trajectories are selected or aligned, and how teacher–student compatibility is improved (Li et al., 2026; Yang et al., 2026; Zhan et al., 2026; Liu et al., 2026a; Lei et al., 2025; Cai et al., 2025; Jiang et al., 2025). Its central design object is therefore the quality and distribution of training signals, rather than the low-rank parameter subspace through which a fixed signal is expressed. A complementary parameter-space analysis characterizes the update geometry of full-parameter OPD and reports early update-subspace locking (Shen et al., 2026). Full-parameter tuning retains all coordinate directions of the OPD gradient, whereas a zero-output low-rank adapter initially retains only the component selected by its row space. Thus, early locking motivates a channel-level question, but does not determine how that channel should be initialized under a rank constraint. We keep the OPD objective unchanged and study the initial route by which its dense teacher signal enters a low-rank update.

Low-rank adaptation and spectral initialization. LoRA parameterizes updates with trainable low-rank factors (Hu et al., 2022; Liu et al., 2024a). The resulting literature treats parameter-efficient adaptation capacity, update geometry, and preservation of pretrained structure as central bottlenecks. Orthogonal and spectral methods alter the factor geometry or select principal and minor pretrained components, while task- and direction-aware methods seek subspaces better suited to downstream adaptation (Buyukakyuz, 2024; Liu et al., 2024b; Qiu et al., 2023; Meng et al., 2024; Wang et al., 2025; Lin et al., 2026; Xiao et al., 2026; Guo et al., 2025a). Together, these works show that equal-rank updates are not interchangeable: initialization and subspace choice can preserve different pretrained structure and expose different adaptation directions. Their design target, however, is generic efficient or task-aware fine-tuning rather than the initial projection of a student-conditioned, token-level OPD gradient. Standard LoRA does not deliberately align its initial row space with a pretrained spectral channel, leaving open whether making this route explicit changes what low-rank OPD can retain from the same supervision. Our comparison therefore does not assume that either principal or minor directions are generally preferable; it isolates the consequences of contrasting equal-budget spectral channels in OPD.

Closest orthonormal-initialization work. The closest work studies zero-output orthonormal adapters initialized from principal or minor right-singular directions (Zhang et al., 2026). We reuse this construction—a zero output factor, a row-orthonormal adapter subspace, and pretrained principal or minor right-singular directions—because it controls the initial function, rank, and projection normalization while varying the selected channel. The distinction is the question posed and the evidence required: rather than treating the construction only as a structured initialization for adaptation, we use it as a controlled instrument for low-rank OPD. Specifically, we ask whether the two equal-rank channels route the unchanged OPD gradient differently and whether those differences co-occur with changed task outcomes. The matched-rank benchmark comparisons and trajectory- and spectral-channel analyses are intended to evaluate this OPD-specific design dimension, not to establish a universal ordering among spectral regions.

Parameter-space diagnostics. We use established sparsity, spectral, rotation, and localization diagnostics (Liu et al., 2026b; Mukherjee et al., 2025; Zhu et al., 2025; Shen et al., 2026) to characterize which update structure each parameterization retains. These measurements complement the controlled benchmark comparison by testing whether the observed updates remain associated with the designated pretrained channels across training. They are descriptive consistency checks for the

routing interpretation: benchmark comparisons establish the relative outcomes in the evaluated settings, whereas no individual sparsity, alignment, or trajectory statistic is treated as a causal mediator of the gains.

3 METHODOLOGY

3.1 PROBLEM SETUP AND DESIGN GOAL

For a pretrained linear weight $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA parameterizes the adapted weight as

$$W = W_0 + sBA, \qquad A \in \mathbb{R}^{r \times d_{\text{in}}}, \quad B \in \mathbb{R}^{d_{\text{out}} \times r}, \quad s = \frac{\alpha}{r}. \tag{1}$$

Our design goal is to retain the zero-output property of standard LoRA while controlling the subspace through which the initial OPD gradient enters the rank-r update. The OPD loss and training procedure are otherwise unchanged.

3.2 ORTHONORMAL GRADIENT CHANNEL

Let $G_0 = \nabla_W \mathcal{L}_{\text{OPD}}(W_0)$ denote the initial OPD gradient. We set $B_0 = 0$, so the adapter contributes no change to the initial model function and the gradient with respect to A is zero at initialization. To expose the first-step channel, consider an SGD update to B:

$$B_1 = -\eta s G_0 A_0^\top, \qquad \Delta W_1 = sB_1A_0 = -\eta s^2 G_0 P_0, \qquad P_0 = A_0^\top A_0. \tag{2}$$

Thus, P_0 determines which component of the initial OPD gradient enters the effective weight update. When $A_0 A_0^\top = I_r$, P_0 is a rank-r orthogonal projector. With effective step size $\eta_{\text{eff}} = \eta s^2$,

$$\left\| \frac{\Delta W_1}{\eta_{\text{eff}}} \right\|_F = \|G_0 P_0\|_F \le \|G_0\|_F, \qquad \|G_0 - G_0 P_0\|_F = \|G_0(I - P_0)\|_F. \tag{3}$$

The equations express a standard consequence of LoRA parameterization and row-orthonormality, not a new projection theorem. They isolate the design choice relevant here: orthogonality fixes the projection geometry, while the row space selects which OPD-gradient directions are retained. Our runs use $r = 16$ and $\alpha = 32$, hence $s = 2$; the normalization above is by η_{eff}, and the unnormalized adapter step includes s^2. Since training uses AdamW, this SGD calculation characterizes the initial gradient channel rather than the optimizer's coordinatewise-preconditioned update.

3.3 OP-OPD AND OM-OPD

We instantiate two channels from the pretrained singular value decomposition $W_0 = U\Sigma V^\top$:

$$\begin{aligned} \text{OP-OPD:} \quad & A_0 = V_{[:,1:r]}^\top, \quad B_0 = 0, \\ \text{OM-OPD:} \quad & A_0 = V_{[:,-r:]}^\top, \quad B_0 = 0. \end{aligned} \tag{4}$$

The principal and minor right-singular vectors each give $A_0 A_0^\top = I_r$. Consequently, both variants preserve the initial model function and have the same rank and projection normalization; they differ in the pretrained spectral region selected for the initial gradient channel. In implementation, we compute the SVD in fp32 for each target linear module, copy the selected vectors into A, set B to zero, and then run the unchanged OPD training objective. The complete configurations are listed in Appendix B.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Our primary comparison evaluates standard LoRA-OPD against the proposed OP-OPD and OM-OPD under the same OPD objective. We also report Full OPD as a full-parameter reference. The primary controlled comparison matches the rank and training budget of the three low-rank methods.

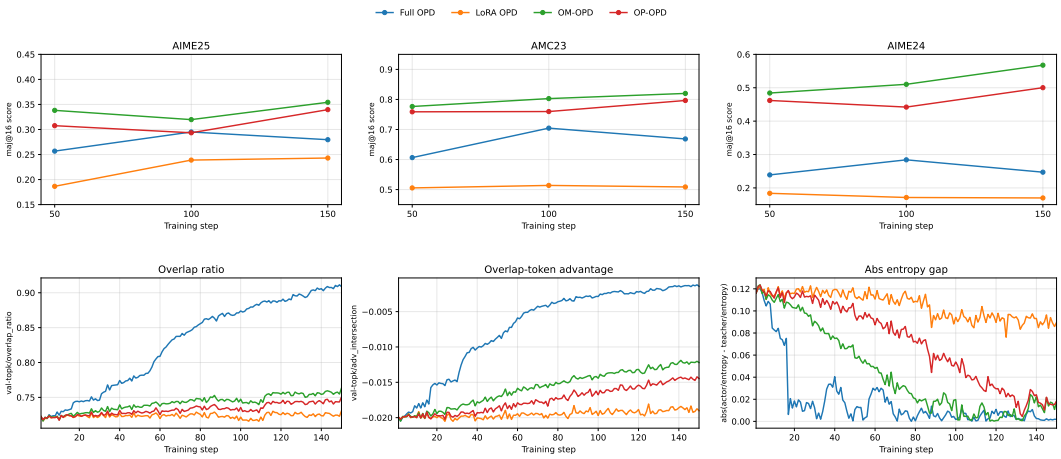

Figure 1: **Primary-setting trajectories.** Top: benchmark maj@16; bottom: token-level signals. Spectral initializations improve trajectories, narrow entropy gaps, and retain comparable top-k overlap.

The two teacher–student pairs reported in the main results are JustRL-DeepSeek-1.5B → DeepSeek-R1-Distill-Qwen-1.5B and JustRL-Nemotron-1.5B → OpenMath-Nemotron-1.5B. Both use DAPO-Math-17k prompts. All low-rank methods target the linear modules with $r = 16$ and $\alpha = 32$; training runs for 150 steps with batch size 64 and learning rate 10^{-6}. The appendix gives the full shared training and validation configuration.

We evaluate mathematical reasoning on AIME25, AMC23, and AIME24. Each problem is sampled 16 times, and we report mean@16, best@16, maj@16, and worst@16. Table 1 reports step-150 outcomes and equally weighted averages over the three benchmarks. The validation bootstrap quantifies resampling uncertainty over evaluation examples; it should not be read as variation across training seeds.

5 Main Results

5.1 Benchmark Performance

Table 1 shows that both spectral initializations improve on standard LoRA-OPD in the two reported settings. On J1r1, OP-OPD improves aggregate mean@16 and maj@16 by 0.1337 and 0.2382, respectively, while OM-OPD improves them by 0.1639 and 0.2734. On OpenMath-Nemotron-1.5B, the corresponding improvements are 0.0087 and 0.0198 for OP-OPD and 0.0484 and 0.0376 for OM-OPD.

Figure 1 shows stronger validation trajectories and narrower entropy gaps for both spectral initializations, with comparable top-k overlap.

6 Mechanism Analysis

This section tests whether the proposed initializations produce the channel-specific update patterns predicted by the construction and characterizes how those patterns co-occur with the benchmark differences above. The controlled benchmark comparison establishes the empirical advantage over standard LoRA-OPD; identifying any individual trajectory or alignment statistic as its causal mediator requires additional interventions.

6.1 Early Update Locking in Standard LoRA-OPD

We first compare standard LoRA-OPD with the proposed initializations using update trajectories, and report Full OPD as an additional reference. Figure 2 shows that standard LoRA-OPD remains

Table 1: Step-150 validation results for both teacher–student settings.

Method	Benchmark	mean@16	best@16	maj@16	worst@16
Teacher → *Student:* JustRL-DeepSeek-1.5B → DeepSeek-R1-Distill-Qwen-1.5B					
LoRA-OPD	AIME25	0.1938	0.3536	0.2430	0.0493
	AMC23	0.4977	0.7451	0.5087	0.2378
	AIME24	0.2042	0.5003	0.1699	0.0497
	Avg.	0.2986	0.5330	0.3072	0.1123
OP-OPD	AIME25	**0.2708**	0.4389	0.3396	0.0930
	AMC23	0.6762	0.8905	0.7965	0.4006
	AIME24	0.3500	**0.7186**	0.5002	0.0744
	Avg.	0.4323	0.6827	0.5454	0.1893
OM-OPD	AIME25	**0.2708**	**0.4417**	**0.3543**	**0.0941**
	AMC23	**0.7146**	**0.9224**	**0.8200**	**0.4504**
	AIME24	**0.4021**	0.7150	**0.5676**	**0.1315**
	Avg.	**0.4625**	**0.6930**	**0.5806**	**0.2253**
Full OPD	AIME25	0.2500	0.3930	0.2796	0.0950
	AMC23	0.6514	0.8696	0.6685	0.4178
	AIME24	0.3042	0.6475	0.2471	0.1240
	Avg.	0.4019	0.6367	0.3984	0.2123
Teacher → *Student:* JustRL-Nemotron-1.5B → OpenMath-Nemotron-1.5B					
LoRA-OPD	AIME25	0.4833	0.6863	0.6210	0.1974
	AMC23	0.8313	0.9515	0.8912	0.6318
	AIME24	0.6396	0.8097	0.7340	0.2851
	Avg.	0.6514	0.8158	0.7487	0.3714
OP-OPD	AIME25	0.5188	0.7293	0.6601	0.2405
	AMC23	0.8449	0.9658	0.9021	0.6647
	AIME24	0.6167	0.8277	0.7433	0.2517
	Avg.	0.6601	0.8409	0.7685	0.3856
OM-OPD	AIME25	**0.5604**	**0.7858**	**0.6862**	**0.2466**
	AMC23	**0.8660**	**0.9706**	**0.9204**	**0.6986**
	AIME24	**0.6729**	**0.8295**	**0.7523**	**0.3970**
	Avg.	**0.6998**	**0.8620**	**0.7863**	**0.4474**
Full OPD	AIME25	0.6062	0.8160	0.6941	0.3608
	AMC23	0.8931	0.9723	0.9324	0.7349
	AIME24	0.6833	0.8326	0.7717	0.4231
	Avg.	0.7275	0.8736	0.7994	0.5063

low in stable rank throughout the measured run, while OP-OPD and OM-OPD end with similarly low stable ranks (approximately 2.7–2.8) but accumulate larger update magnitudes. The full-parameter reference begins with higher stable rank and Hill α, with both declining over training. The Hill estimates use rank-dependent tail ranges, so their levels are interpreted as within-method trajectories rather than direct cross-method comparisons.

The endpoint weight-space diagnostics in Figure 3 provide a complementary comparison between the full-parameter reference and standard LoRA-OPD. Across the analyzed matrices, the full-parameter reference exhibits larger maximum principal-angle excursions and normalized spectral shifts, while the representative normalized singular-value profiles remain broadly similar. Together, these measurements characterize the magnitude and location of weight-space change.

Table 2 reports bf16-aware update sparsity for the two teacher–student pairs. Its complement is the visible-update density. The full-parameter reference changes a larger fraction of coordinates than standard LoRA-OPD in both pairs, quantifying the coordinate-level extent of adaptation.

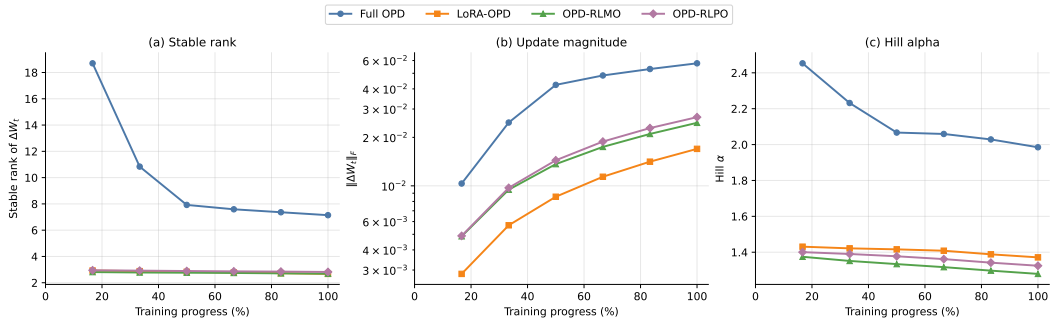

Figure 2: **Intrinsic update trajectories under four OPD parameterizations.** Stable rank, Frobenius update magnitude, and Hill tail estimate over training for the full-parameter reference (Full OPD), LoRA-OPD, OM-OPD, and OP-OPD. The Hill estimator uses rank-dependent tail ranges in this analysis; its cross-method levels are read alongside the stable-rank and update-magnitude trajectories.

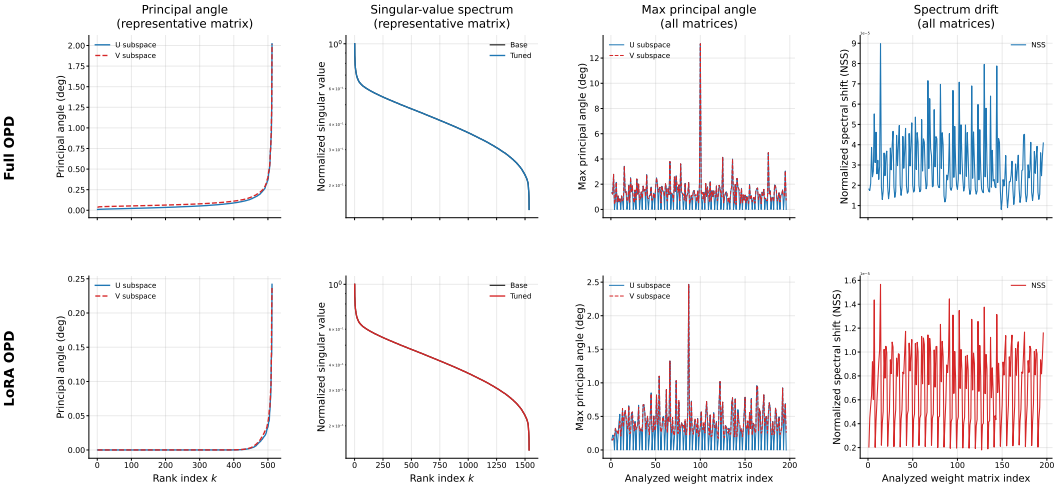

Figure 3: **Weight-space diagnostics for a full-parameter reference and LoRA-OPD.** Rows show Full OPD (top) and LoRA-OPD (bottom). Columns report principal angles for a representative matrix, its normalized singular-value spectrum (base versus tuned), maximum principal angle over analyzed matrices, and normalized spectral shift (NSS) over analyzed matrices. The full-parameter reference shows larger across-matrix geometric excursions, while both parameterizations retain a broadly similar representative singular-value profile.

6.2 Update Geometry Across OPD Parameterizations

Figure 4 shows the endpoint coordinate footprint and conditional overlaps with masks derived from pretrained weights. At step 150, update density is 36.2%, 8.8%, and 12.9% for the full-parameter reference, LoRA-OPD, and OM-OPD, respectively. Their conditional overlap with the principal-weight mask is similar (22.4%, 20.8%, and 20.9%), while low-magnitude-mask overlap is higher for the low-rank runs (68.3%, 99.9%, and 96.3%). OP-OPD is shown in the final row of the figure, completing the coordinate-level view of how the parameterizations allocate their updates.

6.3 Spectral Channel Evidence

Method	A	B	C	D
Sparsity (bf16)	63.9%	91.8%	69.39%	93.76%

Table 2: **bf16-aware update sparsity.** A: DeepSeek-R1-Distill-Qwen-1.5B + JustRL-DeepSeek-1.5B, full-parameter reference (Full OPD); B: DeepSeek-R1-Distill-Qwen-1.5B + JustRL-DeepSeek-1.5B, LoRA-OPD; C: OpenMath-Nemotron-1.5B + JustRL-Nemotron-1.5B, full-parameter reference (Full OPD); D: OpenMath-Nemotron-1.5B + JustRL-Nemotron-1.5B, LoRA-OPD. Visible-update density is the complement of the reported sparsity.

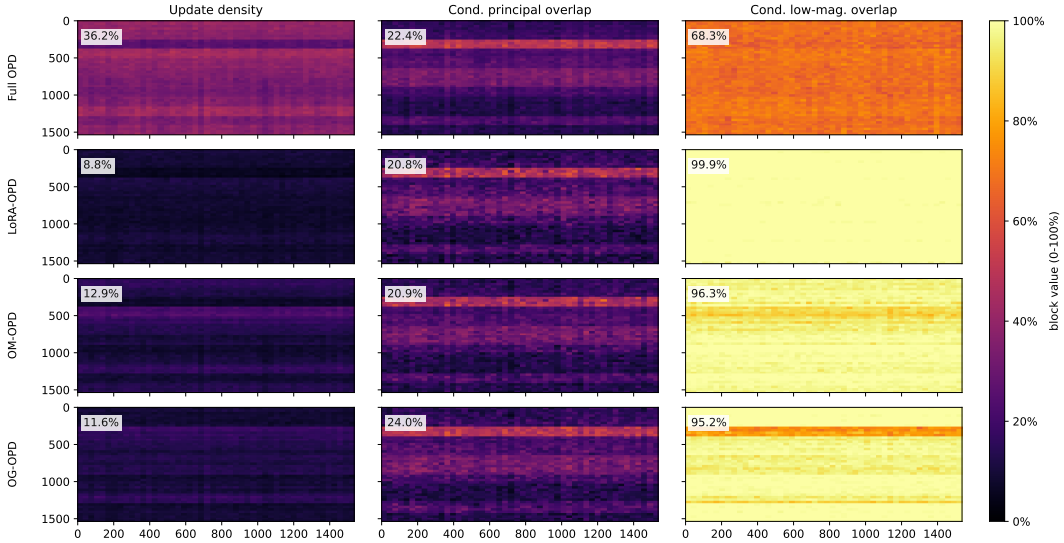

Figure 4: **Update localization at step 150.** Rows correspond to the full-parameter reference (Full OPD), LoRA-OPD, OM-OPD, and OP-OPD. Columns show update density, conditional overlap with the principal-weight mask, and conditional overlap with the low-magnitude mask.

The two proposed initializations select opposite ends of the pretrained right-singular spectrum. Figure 6 shows that OP-OPD's normalized update energy concentrates toward principal modes, whereas OM-OPD's concentrates toward minor modes. Standard LoRA-OPD and the full-parameter reference are broader over the plotted spectrum. Because energy is normalized, this result describes allocation rather than total update magnitude; the latter is shown in Figure 2.

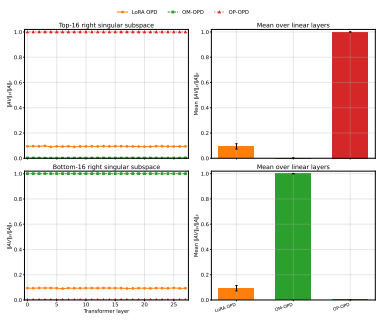

Figure 5: **Adapter spectral alignment.** OP-OPD aligns with principal directions and OM-OPD with minor directions; standard LoRA is low on both.

Figure 5 measures the fraction of each adapter factor captured by the top-16 or bottom-16 pretrained right-singular subspace, $\|AV_F\|_F/\|A\|_F$. OP-OPD aligns with the principal subspace and OM-OPD with the minor subspace across layers; standard LoRA has lower alignment with either extreme. Together, the energy and alignment measurements show that the variants retain distinct associations with their designated spectral channels throughout training, providing the update-space counterpart to the performance improvements in Table 1.

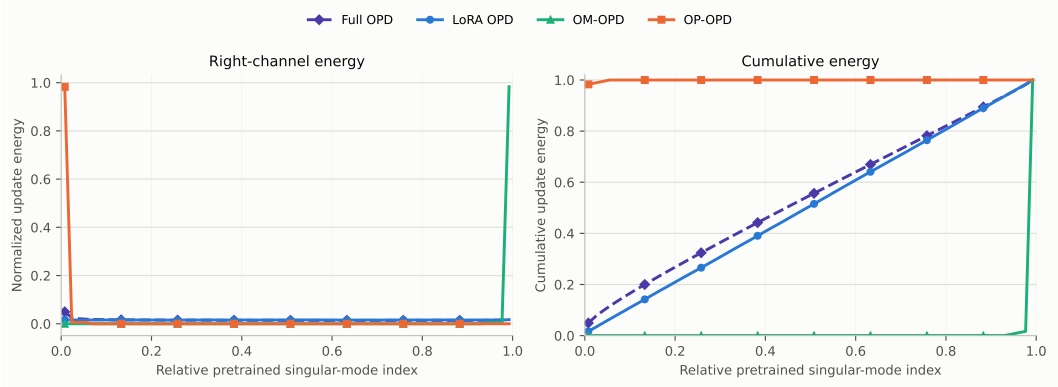

Figure 6: **Normalized update energy across pretrained spectral channels.** The left panel shows energy over relative right-singular-mode index, from principal (0) to minor (1); the right panel shows cumulative energy along the same ordering. OP-OPD concentrates energy toward principal modes, while OM-OPD concentrates it toward minor modes. The plotted distributions describe spectral allocation, not total update magnitude.

7 Conclusion

In this paper, we bridge parameter-efficient fine-tuning and on-policy distillation (OPD) gradient geometry, where standard LoRA yields suboptimal results due to initialization mismatch. Random LoRA prioritizes directions misaligned with OPD's token-level gradient subspace, while ignoring pretrained spectral structure. OP-OPD and OM-OPD address this by selecting principal or minor singular directions via structured SVD, making the update channel a controllable dimension. Evaluations across two teacher-student pairs and three math benchmarks show consistent gains over standard LoRA-OPD, with spectral diagnostics confirming distinct subspace routing. Overall, they provide an effective, stable, and geometrically aligned adaptation strategy for OPD.

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. *arXiv preprint arXiv:2306.13649*, 2024.

Kerim Buyukakyuz. OLoRA: Orthonormal low-rank adaptation of large language models. *arXiv preprint arXiv:2406.01775*, 2024.

Wenrui Cai, Chengyu Wang, Junbing Yan, Jun Huang, and Xiangzhong Fang. Enhancing reasoning abilities of small LLMs with cognitive alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7423–7438. Association for Computational Linguistics, 2025.

Chenlu Guo, Yi Chang, and Yuan Wu. NLoRA: Nyström-initiated low-rank adaptation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1371–1385. Association for Computational Linguistics, 2025a.

Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025b.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

Wangyi Jiang, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. Teach small models to reason by curriculum distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7412–7422. Association for Computational Linguistics, 2025.

Zhenyu Lei, Zhen Tan, Song Wang, Yaochen Zhu, Zihan Chen, Yushun Dong, and Jundong Li. Learning from diverse reasoning paths with routing and collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2832–2845. Association for Computational Linguistics, 2025.

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. *arXiv preprint arXiv:2604.13016*, 2026.

Weicheng Lin, Yi Zhang, Jiawei Dang, and Liang-Jie Zhang. TLoRA: Task-aware low rank adaptation of large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2026.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, 2024a.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, and Songyou Peng. Parameter-efficient orthogonal finetuning via butterfly factorization. In *International Conference on Learning Representations*, 2024b.

Zhenghao Liu, Zhuoyang Wu, Xinze Li, Yukun Yan, Shuo Wang, Zulong Chen, Yu Gu, Ge Yu, and Maosong Sun. Long-chain reasoning distillation via adaptive prefix alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2026a.

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. *arXiv preprint arXiv:2506.00772*, 2026b.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. *arXiv preprint arXiv:2505.11711*, 2025.

Lujie Niu, Haochen Sun, Fangkun Zhao, Sheng Chen, Zimeng Bai, Jiawei Zhang, Caixia Yuan, and Xiaojie Wang. CoTD-PO: Chain-of-thought distillation with preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19975–19986. Association for Computational Linguistics, 2025.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Sch"olkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Zhennan Shen, Yanshu Li, Qingyu Yin, Chak Tou Leong, Zhilin Wang, Yanxu Chen, Rongduo Han, Sunbowen Lee, and Yi R. Fung. On the geometry of on-policy distillation. *arXiv preprint arXiv:2606.07082*, 2026.

Shaowen Wang, Linxin Yu, Jian Li, Xiangyang Liu, and Jun Zhang. MiLoRA: Harnessing minor singular components for parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Xi Xiao, Chenrui Ma, Yunbei Zhang, Chen Liu, Zhuxuanzi Wang, Yanshu Li, Lin Zhao, Guosheng Hu, Tianyang Wang, and Hao Xu. Not all directions matter: Towards structured and task-aware low-rank model adaptation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2026.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, Yunhua Zhou, Yicheng Zou, Qipeng Guo, Tao Gui, Qi Zhang, and Xuanjing Huang. Which reasoning trajectories teach students to reason better? a simple metric of informative alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2026.

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. *arXiv preprint arXiv:2503.14476*, 2025.

Weixiao Zhan, Yongcheng Jing, Leszek Rutkowski, and Dacheng Tao. Distillation traps and guards: A calibration knob for LLM distillability. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2026.

Ruijia Zhang, Jiacheng Zhu, Hanqing Zhu, and Laixi Shi. Geometry-preserving orthonormal initialization for low-rank adaptation in RLVR. *arXiv preprint arXiv:2606.31813*, 2026.

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.

A Statements

AI use statement

Generative AI tools were used to help draft and organize manuscript text, LaTeX structure, and experiment-plan descriptions. The authors are responsible for verifying all claims, equations, experiments, and reported numbers before submission.

Ethics statement

This work studies post-training methods for language models on mathematical reasoning benchmarks. We do not release new human-subject data. The main risks are those generally associated with improving language-model capability; any released checkpoints or code should follow the usage restrictions of the base models and datasets.

Reproducibility statement

The paper specifies the reported teacher–student pairs, training data, validation protocol, LoRA rank and scaling, and the exact initialization formulas for OP-OPD and OM-OPD. Full training and diagnostic details are listed in the appendix and will be accompanied by scripts in the completed version.

Method	Trainable parameters and initialization	Steps	Batch	LR
J1r1: teacher JustRL-DeepSeek-1.5B; student DeepSeek-R1-Distill-Qwen-1.5B; training data DAPO-Math-17k.				
Full OPD	All model parameters	150	64	10^{-6}
LoRA-OPD	All-linear LoRA; $r = 16$, $\alpha = 32$; standard LoRA initialization	150	64	10^{-6}
OM-OPD	All-linear LoRA; $r = 16$, $\alpha = 32$; $A_0 = V_{-r}^\top$, $B_0 = 0$	150	64	10^{-6}
OP-OPD	All-linear LoRA; $r = 16$, $\alpha = 32$; $A_0 = V_r^\top$, $B_0 = 0$	150	64	10^{-6}
J1n1: teacher JustRL-Nemotron-1.5B; student OpenMath-Nemotron-1.5B; training data DAPO-Math-17k.				
Full OPD	All model parameters	150	64	10^{-6}
LoRA-OPD	All-linear LoRA; $r = 16$, $\alpha = 32$; standard LoRA initialization	150	64	10^{-6}
OM-OPD	All-linear LoRA; $r = 16$, $\alpha = 32$; $A_0 = V_{-r}^\top$, $B_0 = 0$	150	64	10^{-6}
OP-OPD	All-linear LoRA; $r = 16$, $\alpha = 32$; $A_0 = V_r^\top$, $B_0 = 0$	150	64	10^{-6}

Table 3: Training configurations from the experiment scripts. Full OPD is the full-parameter reference. V_r and V_{-r} denote the selected principal and minor right-singular directions of the pretrained weight, respectively.

B Experimental Details

Table 3 records the teacher–student configurations and four OPD parameterizations, including Full OPD as the full-parameter reference.

Shared training and evaluation settings. All runs use the token-level direct-reward estimator, four student rollouts per prompt, student and teacher temperatures of 1.0, log-probability top-$k = 16$, the student-only top-k strategy, student-probability reward weighting, rollout top-$p = 1.0$, and repetition penalty 1.0. The training loss uses token-mean aggregation; KL loss and format reward are disabled. Prompts are limited to 1,024 tokens and training responses to 7,168 tokens. Runs use fp32 model dtype on one node with eight GPUs, save every 25 steps, and validate every 50 steps. Validation uses AIME25, AMC23, and AIME24, with 16 sampled responses per problem, temperature 0.7, top-$p = 0.95$, and a 31,744-token response limit. We report mean@16, best@16, maj@16, and worst@16.

B.1 Spectral-Channel Comparison

OP-OPD and OM-OPD share rank, scaling, zero-output initialization, and row-orthonormality; they differ only in whether the initial adapter spans principal or minor pretrained right-singular directions. Table 1 shows that both choices improve on standard LoRA-OPD across the two reported teacher–student pairs. The minor channel yields the larger aggregate improvement in these experiments, while both spectral extremes provide a stronger low-rank initialization than the standard LoRA choice.

C Diagnostic Definitions

Stable rank. For cumulative update ΔW_t, stable rank is

$$\mathrm{srank}(\Delta W_t) = \frac{\|\Delta W_t\|_{\mathrm{F}}^2}{\|\Delta W_t\|_2^2}.$$

Frobenius norm. We report $\|\Delta W_t\|_{\mathrm{F}}$ to measure update scale and to rule out small-update explanations for low stable rank.

Early-final subspace similarity. Let $V_K(t)$ denote the top-K right singular subspace of ΔW_t. We report

$$\mathrm{Sim}_K(t, t_{\mathrm{end}}) = \frac{1}{K}\|V_K(t)^\top V_K(t_{\mathrm{end}})\|_{\mathrm{F}}^2.$$

Normalized spectral shift. For endpoint weight W_+ and initialization W_0,

$$\mathrm{NSS}(W_0, W_+) = \frac{\|\sigma(W_+) - \sigma(W_0)\|_2}{\|\sigma(W_0)\|_2}.$$

Update-mask overlap. We compare a visible update mask with a principal mask derived from a low-rank SVD reconstruction of W_0 and a low-magnitude mask derived from entries of small $|W_0|$. These are threshold-dependent descriptive overlaps; the analysis script's rank, threshold, and aggregation settings must accompany any quantitative claim.

D Full Validation Curves

Table 4: Full validation curves for mean@16, including the full-parameter reference (Full OPD).

Method	Step	AIME25	AMC23	AIME24	Avg.
Full OPD	50	0.2333	0.5964	0.2687	0.3661
Full OPD	100	0.2562	0.6627	0.3146	0.4112
Full OPD	150	0.2500	0.6514	0.3042	0.4019
LoRA-OPD	50	0.1729	0.5098	0.2167	0.2998
LoRA-OPD	100	0.1896	0.5060	0.2021	0.2992
LoRA-OPD	150	0.1938	0.4977	0.2042	0.2986
OP-OPD	50	0.2250	0.6401	0.3208	0.3953
OP-OPD	100	0.2271	0.6551	0.3146	0.3989
OP-OPD	150	0.2708	0.6762	0.3500	0.4323
OM-OPD	50	0.2583	0.6581	0.3229	0.4131
OM-OPD	100	0.2542	0.6867	0.3625	0.4345
OM-OPD	150	0.2708	0.7146	0.4021	0.4625