GeoRA: Geometry-Aware Low-Rank Adaptation for RLVR

Jiaying Zhang[1,2†*], **Lei Shi**[1*‡], **Jiguo Li**[1],
Jun Xu[1‡], **Jiuchong Gao**[1‡], **Jinghua Hao**[1], **Renqing He**[1]
[1]Meituan, [2]Peking University
zhangjy2002@stu.pku.edu.cn
{shilei74, xujun58, gaojiuchong}@meituan.com

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) is a key paradigm for improving large-scale reasoning models. Unlike supervised fine-tuning (SFT), RLVR exhibits distinct optimization dynamics and is sensitive to the preservation of pre-trained geometric structures. However, existing parameter-efficient methods face key limitations in this regime. Low-rank adaptation methods, such as PiSSA, are primarily designed for Supervised Fine-Tuning (SFT) and do not account for the distinct optimization dynamics and geometric structures of RLVR. Conversely, directly fine-tuning the unstructured sparse parameter subspace favored by RLVR encounters efficiency bottlenecks on modern hardware. To address these challenges, we propose GeoRA (Geometry-Aware Low-Rank Adaptation), a low-rank adaptation method tailored for RLVR. Specifically, GeoRA exploits the anisotropic and compressible structure of RL update subspace, and extracts its principal directions via Singular Value Decomposition (SVD) to initialize low-rank adapters, while freezing residual components as a structural anchor during training. This design preserves the pre-trained structure and enables efficient dense computation. Experiments on Qwen and Llama models from 1.5B to 32B parameters show that GeoRA consistently outperforms strong low-rank baselines across RLVR settings in mathematics, medicine, and coding, while showing stronger generalization and less forgetting on out-of-domain tasks.

1 Introduction

Large reasoning models, represented by OpenAI-o1 (OpenAI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025), have established Reinforcement Learning with Verifiable Rewards (RLVR) as a pivotal paradigm for unlocking complex reasoning capabilities. Unlike supervised fine-tuning (SFT), RLVR is better characterized as a constrained optimization process (Wu et al., 2025) that amplifies latent reasoning behaviors through reward-induced sampling bias (Yue et al., 2025; Zhao et al., 2025). As a result, RLVR is particularly sensitive to update stability and its alignment with pre-trained representation geometry: overly aggressive updates can collapse behavior or degrade general capabilities. Empirically, substantial gains can emerge from modifying only a small fraction of parameters (Mukherjee et al., 2025), and recent mechanistic studies further suggest that effective RLVR updates are geometrically biased toward preserving pre-trained structure (Zhu et al., 2025; Cai et al., 2025).

However, existing PEFT methods face key limitations in this regime. First, SFT-oriented low-rank adaptation methods suffer from a geometric mismatch with RLVR (Yin et al., 2025). PiSSA (Meng et al., 2025), for example, forces updates onto principal components, conflicting with RLVR's preferred subspace while protecting core features. Second, some sparse fine-tuning methods (Mukherjee et al., 2025; Zhu et al., 2025), although more consistent with RLVR update patterns, struggle to achieve practical efficiency. Because modern hardware provides limited support for unstructured sparsity, these methods often fail to translate theoretical sparsity into real-world speedups, and may even introduce additional overhead.

To address these challenges, we introduce **GeoRA** (Geometry-Aware Low-Rank Adaptation), a low-rank adaptation framework tailored to RLVR. Our analysis shows that the effective RLVR update subspace is anisotropic and compressible rather than isotropic, exhibiting a structured low-rank form. Based on this observation, GeoRA extracts dominant trainable directions from a geometry-constrained subspace for initialization, while keep-

† Work done during internship at Meituan.
‡ Corresponding authors.
* Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24207–24221

ing residual components frozen as a structural anchor. In this way, GeoRA simultaneously aligns adaptation with RLVR-specific optimization geometry and preserves dense matrix computation, achieving both stable optimization and hardware-efficient training. To the best of our knowledge, GeoRA is the first geometry-aware low-rank adaptation framework explicitly designed for RLVR. Our contributions are summarized as follows:

- We propose GeoRA, a geometry-aware, low-rank, and parameter-efficient adaptation framework tailored to RLVR. By aligning low-rank adaptation with RLVR-specific optimization geometry while preserving dense computation, GeoRA overcomes the geometric mismatch of SFT-oriented low-rank methods and the efficiency bottleneck of sparse methods.
- We show that the effective RL update subspace is directional and admits a compressible low-rank structure. GeoRA extracts dominant trainable directions via SVD within this subspace to initialize low-rank adapters, while a frozen residual component acts as a structural anchor to preserve pre-trained structure.
- Experiments on Qwen and Llama models from 1.5B to 32B parameters demonstrate that GeoRA improves training stability and consistently outperforms strong low-rank baselines across diverse RLVR settings, while showing stronger generalization and less forgetting on out-of-domain tasks.

2 Related Work

2.1 RLVR and Optimization Geometry

RLVR replaces traditional reward models with deterministic verifiers (e.g., in math or coding) (Zhang et al., 2025; Lambert et al., 2025; Yuan et al., 2025). Through outcome-based feedback, it incentivizes emergent reasoning behaviors like Chain-of-Thought (CoT), establishing itself as a core paradigm for enhancing LLM reasoning (Hu et al., 2025; Liu et al., 2025b).

Recent mechanistic analyses have delineated a sharp dichotomy between supervised fine-tuning (SFT) and reinforcement learning with verifiable rewards (RLVR). While SFT primarily injects knowledge by modifying principal weight directions (Chu et al., 2025; Jin et al., 2025b), RLVR is better characterized as a constrained optimization process (Wu et al., 2025) that amplifies latent reasoning behaviors via reward-induced sampling bias rather than introducing new capabilities (Yue et al., 2025; Zhao et al., 2025; Alam and Rastogi, 2025). Theoretically, these updates manifest "off the principals," favoring low-magnitude directions orthogonal to pre-trained features (Zhu et al., 2025), consistent with the rank-1 dominance observed in early training (Cai et al., 2025). However, this regime faces stability trade-offs. KL-regularization ("RL's Razor") attempts to limit forgetting (Shenfeld et al., 2025) but can precipitate the "Reasoning Boundary Paradox," where aggressive reward maximization collapses exploration diversity (Nguyen et al., 2025). Although capability degradation may be partially reversed via singular vector rotation (Jin et al., 2025a,b), adaptation remains fundamentally constrained by the "Invisible Leash," which enforces proximity to the pre-training manifold (Wu et al., 2025). Empirically, strong gains can emerge from updating only a small fraction of parameters, suggesting that RL fine-tuning often concentrates on small subnetworks (Mukherjee et al., 2025).

2.2 PEFT and Spectral Priors

To alleviate the computational demands of scaling LLMs, PEFT (Xu et al., 2023; Han et al., 2024) has emerged as a key paradigm, minimizing memory overhead by updating only a fraction of parameters while matching full fine-tuning performance. Prevailing strategies include partial fine-tuning (Zaken et al., 2022; Lawton et al., 2023), soft prompt tuning (Hambardzumyan et al., 2021), non-linear adapters (Lin et al., 2020), low-rank adaptation (Li et al., 2018), and importance-aware methods like LIFT (Liu et al., 2025c). As a practical paradigm, PEFT has also been effectively adapted to continual learning (Liang et al., 2023) and multi-task learning (Liu et al., 2023) scenarios.

Among these, LoRA (Hu et al., 2021) and its variants are the de facto standard, yet their initialization strategies often diverge based on spectral priors. For instance, PiSSA (Meng et al., 2025) allocates trainable parameters to principal singular components, imposing a strong inductive bias validated primarily in SFT. However, such SFT-oriented spectral priors can create a fundamental geometric mismatch in RLVR, whose optimization dynamics and effective update patterns differ markedly from SFT (Yin et al., 2025; Zhu et al., 2025; Mukherjee et al., 2025). MiLoRA (Wang et al., 2025) instead targets minor components,

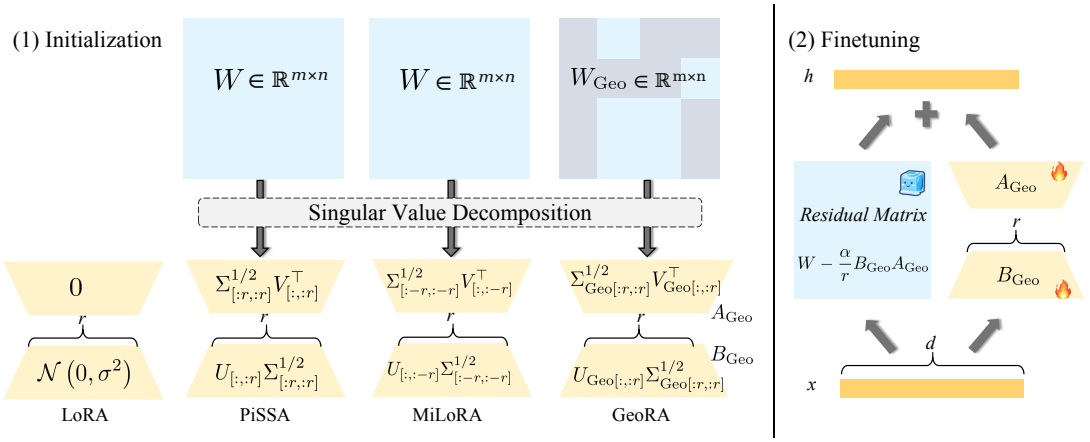


Figure 1: Comparison of adapter initialization and forward architectures. **LoRA** applies low-rank adaptation on the original weight matrix W with standard initialization, while **PiSSA** initializes adapters from the principal components of W. In contrast, **GeoRA** initializes from a geometry-constrained matrix W_{Geo} (a different adaptation target than W). Its forward pass incorporates a frozen Residual Matrix in parallel with the trainable adapter to act as a stability anchor for principal components.

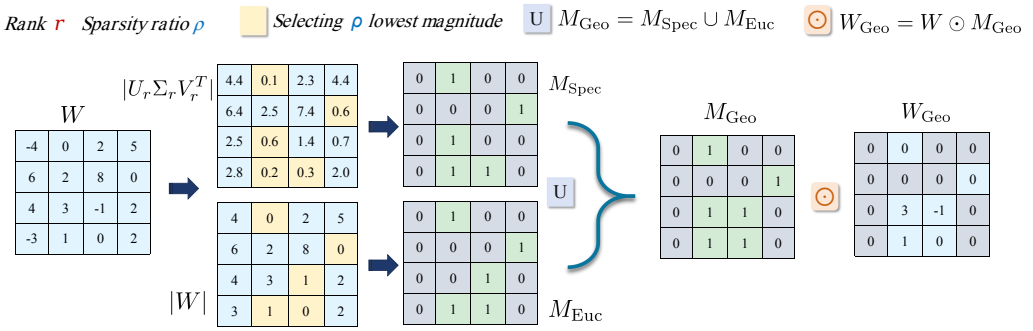


Figure 2: Geometric Prior Construction via Masking. The process of generating M_{Geo} by combining Spectral Priors (low-curvature regions) and Euclidean Priors (high-plasticity near-zero weights). The resulting W_{Geo} isolates the most stable parameters for RL-native adaptation.

but does not explicitly account for RLVR-specific geometry and dynamics. Other variants like DoRA (Liu et al., 2024), AdaLoRA (Zhang et al., 2023), and VeRA (Kopiczko et al., 2024) focus on weight decomposition or budget allocation without integrating these RLVR-native geometric considerations. Building on these observations, GeoRA translates RLVR-specific mechanistic insights into an actionable PEFT paradigm.

3 Methodology

We introduce **GeoRA** (Geometry-Aware Low-Rank Adaptation), a low-rank adaptation framework tailored for RLVR. GeoRA emphasizes an RL-native, structured low-rank parameterization together with an explicit residual leash for stability. Comparing with classical PEFT baselines like LoRA and PiSSA, we first present the low-rank forward architecture and initialization (Figure 1), and then describe a lightweight masking instantiation that constructs the geometry-constrained matrix used for initialization (Figure 2).

3.1 Geometry-Aware Low-Rank Structure

Let W_{Geo} denote a geometry-constrained view of the pre-trained weight matrix W. Unlike standard LoRA which initializes adapters randomly (or with zero), GeoRA leverages the compressible structure within W_{Geo} to derive a structured initialization.

We first perform Singular Value Decomposition (SVD) on the geometry-constrained matrix:

$$W_{\text{Geo}} = U_{\text{Geo}} \Sigma_{\text{Geo}} V_{\text{Geo}}^\top \tag{1}$$

We extract the top-r singular components that capture the principal geometric information. The low-rank adapters A_{Geo} and B_{Geo} are then initialized to approximate this geometry-aware subspace:

$$A_{\text{Geo}} = \Sigma^{1/2}_{\text{Geo}[:r,:r]} V^{\top}_{\text{Geo}[:,:r]} \tag{2}$$

$$B_{\text{Geo}} = U_{\text{Geo}[:,:r]} \Sigma^{1/2}_{\text{Geo}[:r,:r]} \tag{3}$$

By this design, the initial product $B_{\text{Geo}} A_{\text{Geo}}$ constructs the rank-r approximation of W_{Geo}.

Crucially, to ensure the model's output remains unchanged at initialization and to preserve core capabilities during training, we follow the standard LoRA scaling and compute a Residual Matrix W_{res} by subtracting the scaled initialized adapters from the original weights:

$$W_{\text{res}} = W - \frac{\alpha}{r} B_{\text{Geo}} A_{\text{Geo}} \tag{4}$$

During the forward pass, W_{res} is kept **frozen**. The hidden state h is computed as:

$$h = \underbrace{W_{\text{res}}}_{\text{Frozen}} x + \underbrace{\frac{\alpha}{r} B_{\text{Geo}} A_{\text{Geo}}}_{\text{Trainable}} x \tag{5}$$

This construction keeps the model function-preserving at initialization (since $W_{\text{res}} x + \frac{\alpha}{r} B_{\text{Geo}} A_{\text{Geo}} x = Wx$) and enforces a hard structural constraint: the optimizer can only update the geometry-aligned manifold parameterized by A_{Geo} and B_{Geo}, while W_{res} acts as a stability anchor preventing the erosion of pre-trained representations.

3.2 Geometric Prior Construction

To instantiate a geometry-aware update region, we construct a geometry-constrained matrix W_{Geo} using a masking strategy. This masking is consistent with prior observations on spectral and magnitude structure and stability in fine-tuning (Liu et al., 2025c; Zhu et al., 2025).

The Spectral Prior (M_{Spec}) promotes geometric stability by selecting the bottom ρ-fraction of entries from the rank-r approximation $\hat{W}_r$. The mask is defined as:

$$(M_{\text{Spec}})_{i,j} = I\left(|(\hat{W}_r)_{i,j}| \leq \tau_{\text{Spec}}(\rho)\right) \tag{6}$$

where $\tau_{\text{Spec}}(\rho)$ is the ρ-th quantile of the absolute values in $\hat{W}_r$. Intuitively, this mask suppresses high-magnitude (and typically high-curvature) components and constrains updates to a more stable, low-magnitude region, improving spectral stability under RLVR. Similarly, the Euclidean Prior (M_{Euc}) selects low-magnitude weights to capture parameter plasticity, using the same sparsity ratio ρ:

$$(M_{\text{Euc}})_{i,j} = I\left(|W_{i,j}| \leq \tau_{\text{Euc}}(\rho)\right) \tag{7}$$

Here, $\tau_{\text{Euc}}(\rho)$ represents the ρ-th quantile of $|W|$. The final geometry-constrained matrix W_{Geo} is formed by the **union** of these two stable subspaces:

$$W_{\text{Geo}} = W \odot (M_{\text{Spec}} \cup M_{\text{Euc}}) \tag{8}$$

This union ensures that the optimized weights retain the flexibility of small parameters while respecting the spectral constraints of the pre-trained model.

4 Experiments

4.1 Experimental Setup

We compare GeoRA against MiLoRA (Wang et al., 2025), PiSSA (Meng et al., 2025), LoRA (Hu et al., 2021), Sparse Fine-Tuning (SparseFT) (Zhu et al., 2025), and Full Fine-Tuning (FullFT). Our main experiments are conducted on mathematical RLVR, where we fine-tune Qwen3-8B-Base (Yang et al., 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on the DeepMath-103K dataset (He et al., 2025) using the GRPO algorithm (Shao et al., 2024), with a fixed rank $r = 16$ and sparsity ratio $\rho = 0.2$. We include both base and instruction-tuned backbones to evaluate robustness across model variants. In addition, we further validate our method on mathematical RLVR tasks with the 4B and 1.5B models on the GSM8K (Cobbe et al., 2021) dataset; details are provided in the Appendix C.

For in-distribution (ID) evaluation, we assess mathematical reasoning performance on AIME24, AIME25 (Patel et al., 2024), MATH-500 (Hendrycks et al., 2021b), and OlymMATH (Sun et al., 2025). For out-of-distribution (OOD) evaluation, we consider HumanEval (Coding) (Chen et al., 2021), GPQA (Science) (Rein et al., 2023), MMLU (General Knowledge) (Hendrycks et al., 2021a), IFEval (Instruction Following) (Zhou et al., 2023), and TruthfulQA (Truthfulness) (Lin et al., 2022) to measure cross-domain generalization and capability preservation. We also report the pre-RLVR Base model performance for direct comparison.

Beyond the main mathematical setting, we further evaluate GeoRA on additional RLVR domains, including medical reasoning and coding. For medical RLVR, we train Llama-3.1-8B-Instruct on AlphaMed19K (Liu et al., 2025a) and report results

Table 1: Main results on in-distribution (ID) mathematical RLVR benchmarks and out-of-distribution (OOD) tasks. Base denotes the original model before RLVR. Best results are in bold, and second-best results are underlined.

	In-Distribution (ID)				Out-of-Distribution (OOD)				
Method	AIME24	AIME25	MATH500	OlymMATH	HumanEval	GPQA	MMLU	IFEval	TruthfulQA
Qwen3-8B									
Base	13.33	11.67	71.20	9.75	76.83	36.91	71.94	**54.32**	**68.91**
FullFT	23.33	**22.08**	**78.40**	11.25	76.83	36.91	71.94	50.45	65.65
SparseFT	22.92	21.25	76.80	11.50	79.50	37.20	74.20	50.95	66.05
LoRA	19.58	19.58	75.60	10.75	81.10	37.50	75.65	52.13	66.82
PiSSA	22.50	20.42	74.40	11.75	71.95	36.16	73.89	48.74	65.95
MiLoRA	20.42	19.58	76.20	11.50	78.66	**38.26**	74.51	51.85	66.46
GeoRA	**23.75**	21.67	78.00	**12.75**	**82.93**	37.92	**75.96**	53.73	68.85
Llama-3.1-8B									
Base	9.58	2.08	51.00	3.25	65.20	31.15	68.40	**79.99**	**62.71**
FullFT	18.33	8.25	**62.40**	8.50	65.20	31.15	68.40	75.92	59.40
SparseFT	17.92	8.10	61.50	8.25	67.80	31.65	69.10	76.61	60.19
LoRA	15.42	6.25	58.20	6.75	69.80	32.10	69.80	77.93	61.76
PiSSA	17.50	7.92	60.50	7.75	67.50	31.80	69.20	74.83	59.83
MiLoRA	16.25	7.08	59.10	7.25	68.20	32.00	70.50	78.31	61.47
GeoRA	**18.54**	**8.75**	61.90	**8.85**	**70.80**	**32.65**	**70.95**	78.72	61.61

Table 2: Results on additional medical and coding RLVR tasks. Best results are in bold, and second-best results are underlined.

	Medical				Coding			
Method	MedQA	MedMCQA	PubMedQA	Average	LiveCodeBench	HumanEval	MBPP	Average
Base	58.03	56.42	74.94	63.13	65.75	87.66	79.40	77.60
FullFT	75.32	**64.56**	80.12	73.33	**67.75**	**90.26**	81.40	**79.80**
LoRA	74.23	62.12	79.54	71.96	67.25	88.96	81.00	79.07
GeoRA	**76.12**	64.31	**80.64**	**73.69**	**67.75**	89.61	**81.60**	79.65

on MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019). For coding, we train Qwen3-32B on Eurus-2-RL-Data (Cui et al., 2025) and evaluate on LiveCodeBench (Jain et al., 2025), HumanEval, and MBPP (Austin et al., 2021). The implementation details and hyperparameter settings are provided in the Appendix C.

4.2 Main Results

Mathematical Reasoning Performance. Table 1 shows that GeoRA achieves the strongest overall performance on mathematical RLVR benchmarks across both backbones. It obtains the best results on most AIME and OlymMATH settings, while remaining competitive on MATH500. In particular, GeoRA consistently outperforms LoRA, MiLoRA, and PiSSA on challenging competition-style benchmarks, suggesting that geometry-aware initialization provides a more suitable inductive bias for RLVR than classic low-rank initialization.

Out-of-Distribution Performance. GeoRA also generalizes favorably to OOD tasks. Across both backbones, it achieves the strongest or near-strongest results on HumanEval, GPQA, and MMLU, while generally retaining higher capability than other fine-tuning baselines on IFEval and TruthfulQA. These results indicate that constraining updates to geometry-aligned subspaces can improve in-domain reasoning performance while reducing interference with pre-existing general capabilities.

Extension Beyond Mathematical RLVR. Beyond the main mathematical RLVR setting, we further evaluate GeoRA on medical and coding RLVR tasks. As shown in Table 2, GeoRA consistently outperforms LoRA across both domains and remains competitive with FullFT, suggesting that its advantage is not limited to mathematical reasoning.

4.3 Training Stability and Efficiency

Training Dynamics. Figure 3 shows the training dynamics of Qwen3-8B as evaluated on the AIME

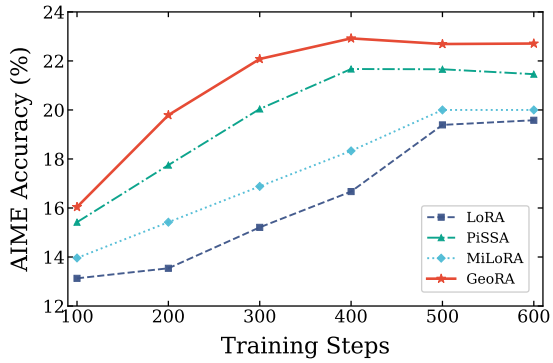


Figure 3: Training dynamics of Qwen3-8B as evaluated on the AIME benchmark (average of 2024 and 2025). GeoRA remains consistently top-performing throughout training.

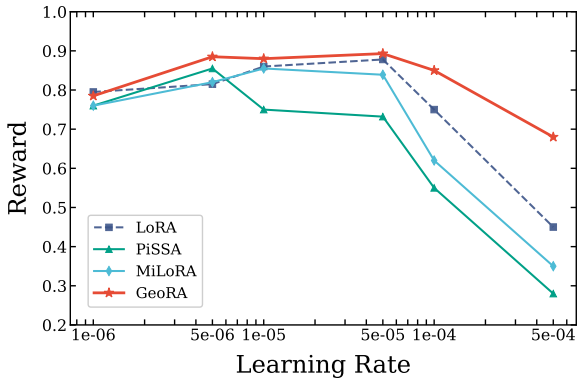


Figure 4: Performance comparison across different learning rates. GeoRA demonstrates superior stability and robust convergence even at higher learning rates.

benchmark. GeoRA remains consistently top-performing throughout training and reaches strong performance substantially earlier than other low-rank baselines. In contrast, LoRA, MiLoRA, and PiSSA improve more slowly and plateau at lower levels. This indicates that GeoRA not only yields better final accuracy, but also provides a more favorable optimization trajectory under RLVR.

Hyperparameter Robustness. We first evaluate performance under different learning rates on Qwen3-4B in Figure 4. GeoRA maintains high reward across a broad range of learning rates, while other low-rank baselines are much more sensitive to this choice. In particular, PiSSA and MiLoRA degrade rapidly under larger learning rates, and LoRA also exhibits a clear drop in performance at the high end of the sweep. These results suggest that GeoRA is more robust to hyperparameter variation and requires less delicate tuning in practice. Similar robustness trends are also observed for other hyperparameters, including rank and sparsity, as shown in Appendix D.

Stability Under Aggressive Optimization. To stress-test optimization stability, we further compare reward and KL divergence under an aggressive learning rate in Figure 5. GeoRA maintains the highest reward trajectory without collapse, whereas PiSSA suffers a catastrophic drop late in training. At the same time, GeoRA keeps KL divergence at a low and controlled level throughout training. These results are highly relevant to RLVR optimization, which often requires stable policy updates within trust-region-like boundaries. The combination of higher reward and smoother KL behavior suggests that GeoRA improves exploration and policy refinement without inducing the destabilizing updates caused by geometry-misaligned adaptation.

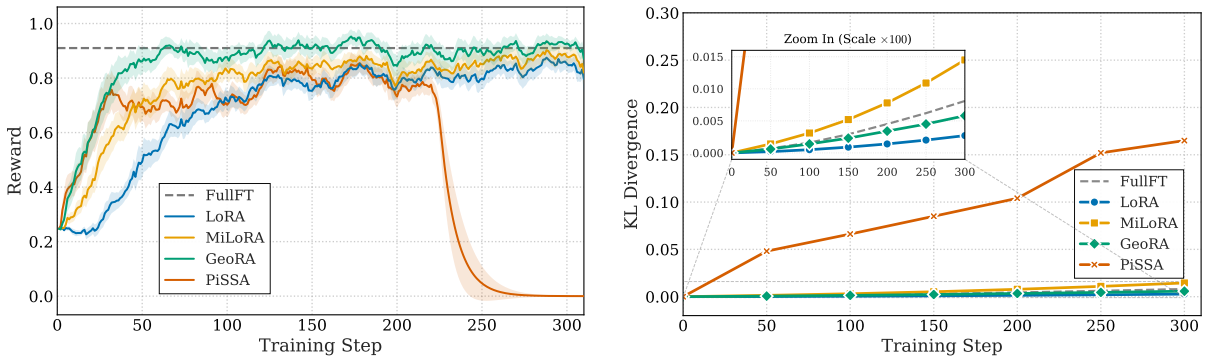
Efficiency. Table 3 compares the training efficiency of FullFT, SparseFT, and GeoRA. GeoRA updates only a tiny fraction of parameters, reducing trainable parameters by 99.5% relative to FullFT, while also lowering VRAM usage and improving per-iteration training speed. Compared with SparseFT, GeoRA avoids the overhead of unstructured sparse computation and therefore translates its parameter efficiency into actual hardware efficiency. Although GeoRA introduces an SVD-based initialization step, this cost is a one-time preprocessing overhead and is negligible compared with RLVR training; detailed profiling is provided in the appendix. Overall, these results show that GeoRA improves not only effectiveness and stability, but also practical training efficiency.

Table 3: Efficiency Comparison between Full FT, SparseFT, and GeoRA. Relative reductions compared to Full FT are shown in parentheses.

Method	Params (B)	Time (s/it)	VRAM (%)
Full FT	8.00	231	95.73
SparseFT	2.56 (-68.0%)	256 (+10.8%)	81.25 (-15.1%)
GeoRA	**0.04** (-99.5%)	**185** (-19.9%)	**68.43** (-28.5%)

4.4 Ablation and Analysis

Impact of Initialization Strategy. Table 4 shows that GeoRA's geometry-aware initialization is critical to its performance. Replacing the proposed initialization with Random-r initialization leads to a clear drop across all benchmarks, while Tail-r initialization performs even worse. This indi-



(a) Reward Trajectories

(b) KL Divergence

Figure 5: Training Stability and Constraint Adherence. Results on Qwen3-4B show that GeoRA demonstrates superior robustness under aggressive learning rates (5×10^{-5}).

Method	Reward	AIME24	AIME25	MATH-500	OlymMATH	Average
GeoRA	**0.88**	**13.33**	**9.17**	**73.40**	**5.75**	**25.41**
Ablation on Initialization Strategy						
Random-r Init	0.85	12.50	8.50	72.10	5.25	24.60
Tail-r Init	0.82	11.67	7.50	70.80	4.50	23.40
Ablation on Geometric Masks						
w/o M_{Spec}	0.86	12.50	8.33	72.00	4.75	24.40
w/o M_{Euc}	0.83	13.33	8.75	72.80	5.50	25.10

Table 4: Ablation study on Qwen3-4B with different initialization strategies and geometric mask variants.

cates that the benefit of GeoRA does not come from low-rank adaptation alone, but from aligning the initialization with the effective RLVR update subspace. In particular, the poor performance of Tail-r suggests that simply selecting low-energy directions is insufficient; the trainable directions must still capture the dominant structure within the geometry-constrained subspace.

Impact of Geometric Masks. We further ablate the two mask components used to construct the geometry-constrained subspace. Removing either M_{Spec} or M_{Euc} consistently degrades performance, confirming that both are necessary for strong results. This suggests that the two priors play complementary roles: the spectral mask helps suppress unstable high-energy directions, while the Euclidean mask preserves adaptation flexibility in low-magnitude regions. Their combination therefore provides a better approximation to the effective RLVR update manifold than either component alone.

Overall Analysis. Taken together, the ablation results show that GeoRA's gains arise from the full geometry-aware design rather than from any single isolated choice. The geometry-constrained subspace, structured initialization, and frozen residual anchor work together to improve optimization quality while preserving pre-trained structure. This is consistent with the stability and efficiency results in the previous section, and further supports our claim that RLVR benefits from a low-rank parameterization explicitly aligned with its update geometry.

5 Mechanism Analysis

5.1 Geometric Priors

GeoRA constructs its update subspace from two geometric priors: a spectral prior and a Euclidean prior. The motivation is that RLVR updates are highly selective, tending to exploit under-utilized yet plastic regions of the pre-trained model while avoiding large modifications to dominant principal directions. Accordingly, M_{Spec} suppresses high-energy components in the principal subspace, whereas M_{Euc} selects low-magnitude parameters in the original weight space. Their union defines a subspace that is both stable and expressive for RLVR adaptation.

We further verify that these two priors are complementary rather than redundant in Table 5. On

Table 5: Overlap analysis of the spectral prior M_{Spec} and Euclidean prior M_{Euc} on Qwen3-8B with $\rho = 0.2$.

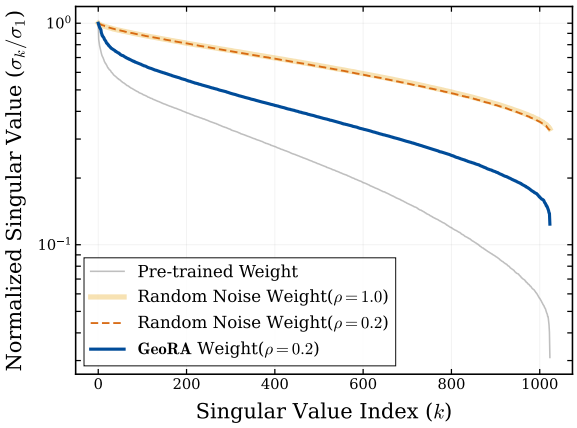
Group	M_{Spec}	M_{Euc}	Intersection	Jaccard
MLP	20.0%	20.0%	4.63%	0.131
Attn	20.0%	20.0%	4.25%	0.119
All	**20.0%**	**20.0%**	**4.55%**	**0.128**

Qwen3-8B with $\rho = 0.2$, both M_{Spec} and M_{Euc} select 20.0% of parameters, but their overlap remains small: the overall intersection is only 4.55%, with a Jaccard index of 0.128. Similar patterns are observed in both MLP layers (4.63%, 0.131) and attention layers (4.25%, 0.119). This limited overlap indicates that the two masks capture largely distinct parameter subsets and therefore serve complementary geometric roles. This is also consistent with the ablation results in Table 4: removing either prior leads to a clear performance drop.

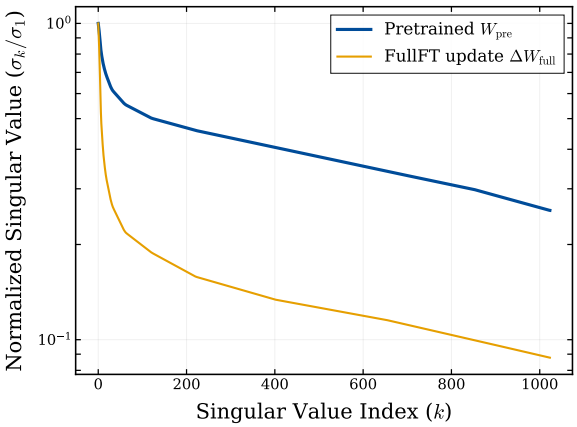
5.2 Low-Rank Structure

To validate the structural basis of GeoRA, we analyze the singular value spectrum of the geometry-constrained subspace. We compare W_{Geo} with the original pre-trained weights, as well as random noise matrices at both full and sparse density. Figure 6a yields two key insights. First, sparsity does not by itself induce low-rankness. The spectrum of sparse random noise is nearly indistinguishable from that of dense random noise, and both follow a relatively flat decay pattern. This indicates that unstructured sparsity alone remains essentially isotropic in the spectral domain. Second, W_{Geo} preserves a pronounced heavy-tailed spectrum similar to that of the pre-trained weights, with most spectral mass concentrated in a small number of leading components. This shows that the selected update region inherits a structured and compressible low-rank form rather than behaving like random sparse noise.

We further find that the actual FullFT update, $\Delta W_{\text{FullFT}} = W_{\text{FullFT}} - W_{\text{Pretrain}}$, exhibits a similarly compressible heavy-tailed spectrum. This strongly supports the low-rank inductive bias of GeoRA: the effective RLVR update space is not isotropic, but already structured and highly compressible. GeoRA therefore captures an intrinsic property of RLVR updates, instead of imposing an artificial low-rank constraint.



(a) Singular value spectrum of the geometry-constrained subspace W_{Geo}.



(b) Singular value spectrum of the full fine-tuning update ΔW_{FullFT}.

Figure 6: Spectral analysis of RLVR update structure.

5.3 Spectral Efficiency and Alignment

Building on the low-rank structure of RL updates, we validate GeoRA's geometric superiority via Normalized Spectral Shift (NSS) and Subspace Alignment(Table 6). First, NSS quantifies the topological distortion of the pre-trained manifold:

$$\text{NSS} = \frac{\|\sigma(W^{\text{tuned}}) - \sigma(W)\|_2}{\|\sigma(W)\|_2}. \tag{9}$$

While PiSSA exhibits high NSS (> 0.39), indicating aggressive structural modification, GeoRA maintains minimal distortion (≈ 0.09), confirming it maximizes reward acquisition while preserving fundamental features.

To pinpoint the update locus, we calculate the alignment between ΔW and the pre-trained singular vectors V:

$$\mathcal{S}(k) = \frac{\|\Delta W v_k\|_2}{\|\Delta W\|_F} \approx |\cos(\theta_k)|. \tag{10}$$

GeoRA shows a distinct signature: it avoids the

Table 6: Geometric mechanism analysis.

	Llama-3.1-8B			Qwen3-8B		
Method	**NSS**↓	$\mathcal{S}_H \downarrow$	$\mathcal{S}_T \uparrow$	**NSS**↓	$\mathcal{S}_H \downarrow$	$\mathcal{S}_T \uparrow$
PiSSA	0.395	0.98	0.01	0.418	0.95	0.03
LoRA	0.214	0.15	0.15	0.235	0.18	0.18
MiLoRA	0.125	0.12	0.92	0.132	0.16	0.90
GeoRA	**0.092**	**0.005**	**0.98**	**0.096**	**0.015**	**0.96**

high-energy principal subspace ($\mathcal{S}_{\text{Head}} \leq 0.02$)—enhancing stability—and efficiently adapts the geometry-constrained tail ($\mathcal{S}_{\text{Tail}} \geq 0.96$). In contrast, PiSSA's instability is explained by its high overlap with the head subspace (≈ 0.98).

6 Conclusion

In this paper, we bridge the gap between parameter-efficient fine-tuning and the distinctive optimization geometry of RLVR. Our study suggests that existing PEFT methods often yield suboptimal results in RLVR due to a structural mismatch. SFT-oriented low-rank methods prioritize directions that may be misaligned with the geometry-constrained updates favored by RLVR, while some sparse update methods better reflect RLVR dynamics but fail to translate into practical hardware efficiency. GeoRA addresses these challenges by identifying a compressible, geometry-aligned update subspace and parameterizing it through structured SVD initialization with a frozen residual anchor. Experiments on Qwen and Llama models from 1.5B to 32B parameters show that GeoRA consistently improves mathematical RLVR performance, extends effectively to medical and coding RLVR settings, and exhibits stronger generalization with less forgetting on out-of-domain tasks. Overall, GeoRA provides an effective, stable, and hardware-efficient adaptation strategy for RLVR.

Limitations

Despite its effectiveness, GeoRA has certain limitations. First, the initialization process requires performing a truncated SVD and dual-masking operations. While these represent a one-time computational cost at the beginning of training, they introduce an additional pre-processing step compared to the random initialization used in standard LoRA. Second, our experiments primarily focus on RLVR in reasoning domains. Although GeoRA demonstrates strong performance on these tasks, its generalizability needs to be further validated across a broader range of model architectures and diverse reinforcement learning scenarios beyond verifiable rewards.

References

Md Tanvirul Alam and Nidhi Rastogi. 2025. Limits of generalization in rlvr: Two case studies in mathematical reasoning. *Preprint*, arXiv:2510.27044.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, and Quoc Le. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Zhenyu Tan, Benyi Zhang, Guiquan Liu, and Junfeng Fang. 2025. On predictability of reinforcement learning dynamics for large language models. *Preprint*, arXiv:2510.00553.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *Preprint*, arXiv:2501.17161.

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

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, and 6 others. 2025. Process reinforcement through implicit rewards. *Preprint*, arXiv:2502.01456.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. Warp: Word-level adversarial reprogramming. *Preprint*, arXiv:2101.00121.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Preprint*, arXiv:2403.14608.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *Preprint*, arXiv:2504.11456.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *Preprint*, arXiv:2503.24290.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nadav Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Hangzhan Jin, Sitao Luan, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. 2025a. Rl fine-tuning heals ood forgetting in sft. *Preprint*, arXiv:2509.12235.

Hangzhan Jin, Sicheng Lv, Sifan Wu, and Mohammad Hamdaqa. 2025b. Rl is neither a panacea nor a mirage: Understanding supervised vs. reinforcement learning fine-tuning for llms. *Preprint*, arXiv:2508.16546.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. 2024. Vera: Vector-based random matrix adaptation. *Preprint*, arXiv:2310.11454.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Neal Lawton, Anoop Kumar, Govind Thattai, Aram Galstyan, and Greg Ver Steeg. 2023. Neural architecture search for parameter-efficient fine-tuning of large pre-trained language models. *Preprint*, arXiv:2305.16597.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. *Preprint*, arXiv:1804.08838.

Zujie Liang, Feng Wei, Yin Jie, Yuxi Qian, Zhenghong Hao, and Bing Han. 2023. Prompts can play lottery tickets well: Achieving lifelong information extraction via lottery prompt tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 277–292.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. *Preprint*, arXiv:2004.03829.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025a. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *Preprint*, arXiv:2505.17952.

Jingping Liu, Tao Chen, Zujie Liang, Haiyun Jiang, Yanghua Xiao, Feng Wei, Yuxi Qian, Zhenghong

Hao, and Bing Han. 2023. Hierarchical prompt tuning for few-shot multi-task learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1556–1565.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025b. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *Preprint*, arXiv:2505.24864.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

Zihang Liu, Tianyu Pang, Oleg Balabanov, Chaoqun Yang, Tianjin Huang, Lu Yin, Yaoqing Yang, and Shiwei Liu. 2025c. Lift the veil for the truth: Principal weights emerge after rank reduction for reasoning-focused supervised fine-tuning. *Preprint*, arXiv:2506.00772.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2025. Pissa: Principal singular values and singular vectors adaptation of large language models. *Preprint*, arXiv:2404.02948.

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

Phuc Minh Nguyen, Chinh D. La, Duy M. H. Nguyen, Nitesh V. Chawla, Binh T. Nguyen, and Khoa D. Doan. 2025. The reasoning boundary paradox: How reinforcement learning constrains language models. *Preprint*, arXiv:2510.02230.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260.

Bhrij Patel, Souradip Chakraborty, Wesley A. Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. 2024. Aime: Ai system optimization via multiple llm evaluators. *Preprint*, arXiv:2410.03131.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark. *Preprint*, arXiv:2311.12022.

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

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2025. Rl's razor: Why online reinforcement learning forgets less. *Preprint*, arXiv:2509.04259.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *Preprint*, arXiv:2503.21380.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. 2025. Milora: Harnessing minor singular components for parameter-efficient llm finetuning. *Preprint*, arXiv:2406.09044.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. 2025. The invisible leash: Why rlvr may or may not escape its origin. *Preprint*, arXiv:2507.14843.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *Preprint*, arXiv:2312.12148.

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qingyu Yin, Yulun Wu, Zhennan Shen, Sunbowen Li, Zhilin Wang, Yanshu Li, Chak Tou Leong, Jiale Kang, and Jinjin Gu. 2025. Evaluating parameter efficient methods for rlvr. *Preprint*, arXiv:2512.23165.

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. 2025. From $f(x)$ and $g(x)$ to $f(g(x))$: Llms learn new skills in rl by composing old ones. *Preprint*, arXiv:2509.25123.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *Preprint*, arXiv:2106.10199.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, and 20 others. 2025. A survey of reinforcement learning for large reasoning models. *Preprint*, arXiv:2509.08827.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *Preprint*, arXiv:2303.10512.

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *Preprint*, arXiv:2504.07912.

Jeffrey Zhou, Tianlu Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

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. 2025. The path not taken: Rlvr provably learns off the principals. *Preprint*, arXiv:2511.08567.

A Algorithm

Algorithm 1 GeoRA Initialization Algorithm

Require: Pre-trained weights $W \in R^{m\times n}$, rank r, sparsity ratio ρ.
Ensure: Initialized adapters A, B, Frozen residual matrix W_{res}.

Phase 1: Geometric Prior Construction
$\hat{W}_r \leftarrow \text{SVD}_r(W)$ {Compute rank-r approximation}
Determine thresholds τ_{spec} and τ_{euc} corresponding to bottom ρ quantile
$M_{\text{spec}} \leftarrow I(|\hat{W}_r| \le \tau_{\text{spec}})$ {Spectral Prior}
$M_{\text{euc}} \leftarrow I(|W| \le \tau_{\text{euc}})$ {Euclidean Prior}
$W_{\text{geo}} \leftarrow W \odot (M_{\text{spec}} \cup M_{\text{euc}})$ {Union of priors (unconstrained sparsity)}
Phase 2: Geometry-Aware Decomposition
$U_g, \Sigma_g, V_g^\top \leftarrow \text{SVD}(W_{\text{geo}})$
$A \leftarrow U_g[:, :r]\Sigma_g^{1/2}$
$B \leftarrow \Sigma_g^{1/2} V_g^\top[:, :r]$
Phase 3: Residual Leash Instantiation
$W_{\text{res}} \leftarrow W - AB$
Freeze W_{res} during training
return A, B, W_{res}

B GeoRA Initialization Properties

This section provides lightweight theoretical justifications for GeoRA's initialization and reparameterization.

B.1 Optimality of the Geo-SVD Initialization

Proposition 1 (Eckart–Young optimality within W_{geo}). Let $W_{\text{geo}} \in R^{m\times n}$ admit an SVD $W_{\text{geo}} = U\Sigma V^\top$, and define the truncated reconstruction $W_{\text{geo}}^{(r)} U_{[:,:r]}\Sigma_{[:r,:r]}V_{[:,:r]}^\top$. Then $W_{\text{geo}}^{(r)}$ is the best rank-r approximation under the Frobenius norm:

$$W_{\text{geo}}^{(r)} \in \arg\min_{\text{rank}(X)\le r} \|W_{\text{geo}} - X\|_F. \quad (11)$$

With the initialization $B \leftarrow U_{[:,:r]}\Sigma_{[:r,:r]}^{1/2}$ and $A \leftarrow \Sigma_{[:r,:r]}^{1/2}V_{[:,:r]}^\top$, we have $BA = W_{\text{geo}}^{(r)}$.

B.2 Initialization and Residual Leash

Proposition 2 (Function-preserving at initialization). Define $W_{\text{res}}W - \frac{\alpha}{r}BA$ and the effective weight $W_{\text{eff}}(A,B)W_{\text{res}} + \frac{\alpha}{r}BA$. Then for any input x, the forward pass satisfies $W_{\text{eff}}(A,B)x = Wx$ at initialization, hence GeoRA is function-preserving at $t = 0$.

Table 7: Empirical initialization cost of GeoRA. We compare standard SVD and randomized SVD across model scales, and report the training cost of Qwen3-8B for reference.

Model	Method	Wall-clock Time (min)	Peak VRAM (GB)
8B	Training	838	∼384
8B	Standard SVD	18.87	16.48
8B	Rand. SVD	**0.21**	**15.78**
14B	Standard SVD	25.78	29.63
14B	Rand. SVD	**0.26**	**28.43**
32B	Standard SVD	35.23	64.05
32B	Rand. SVD	**0.47**	**62.38**
72B	Standard SVD	97.82	146.87
72B	Rand. SVD	**0.72**	**140.41**

Gradient mapping (useful for understanding stability). Let $G\frac{\partial \mathcal{L}}{\partial W_{\text{eff}}}$ denote the gradient w.r.t. the effective weight. By matrix calculus,

$$\frac{\partial \mathcal{L}}{\partial A} = \frac{\alpha}{r}B^\top G, \quad \frac{\partial \mathcal{L}}{\partial B} = \frac{\alpha}{r}GA^\top. \quad (12)$$

Freezing W_{res} constrains learning to the rank-r manifold $\{\frac{\alpha}{r}BA : A \in R^{r\times n}, B \in R^{m\times r}\}$, while the residual leash keeps the model anchored near W.

B.3 Compute and Memory Complexity

The truncated SVD on W_{geo} is a one-off preprocessing step. In practice, one can use truncated or randomized SVD with time roughly linear in r (e.g., $\tilde{O}(mnr)$ for dense operators), which substantially reduces the initialization overhead compared with full SVD.

GeoRA uses dense low-rank updates. Computing $\left(\frac{\alpha}{r}BA\right)x$ can be decomposed into two matrix multiplications, costing $O(rn + mr)$ per layer, which is typically more GPU-friendly than irregular sparse updates at the same parameter budget.

B.4 Empirical Initialization Cost

Practical overhead of SVD initialization. Although GeoRA introduces an additional SVD-based preprocessing step, this cost is incurred only once before training. Since GeoRA only requires the top-r singular components, the initialization can be accelerated with randomized SVD, whose complexity is approximately linear in r for dense operators.

Table 7 reports the wall-clock time and peak VRAM of standard SVD and randomized SVD

Table 8: Comprehensive performance comparison on Qwen2.5-1.5B and Qwen3-4B. We report In-Distribution (ID) mathematical reasoning scores and Out-of-Distribution (OOD) generalization scores.

	In-Distribution (ID)				Out-of-Distribution (OOD)		
Method	AIME24	AIME25	MATH500	OlymMATH	HumanEval	GPQA	MMLU
			Qwen2.5-1.5B				
Full FT	7.92	1.25	53.00	4.00	45.73	25.10	55.40
SparseFT	8.33	1.67	53.60	4.50	51.83	27.50	59.80
LoRA	7.50	0.83	52.60	3.75	**59.15**	29.10	62.10
PiSSA	8.75	1.67	53.40	4.50	57.32	29.50	62.50
MiLoRA	9.58	2.08	**54.80**	5.25	58.54	29.80	**62.80**
GeoRA	**10.83**	**2.50**	54.60	**5.50**	58.54	**30.23**	62.40
			Qwen3-4B				
Full FT	12.92	**9.58**	**73.80**	5.00	3.66	31.46	61.15
SparseFT	12.92	9.17	72.20	5.25	4.27	31.96	63.50
LoRA	10.83	8.33	71.20	4.75	3.05	31.04	62.80
PiSSA	12.50	8.75	70.00	5.25	4.88	32.47	64.95
MiLoRA	11.25	8.33	71.80	5.25	4.88	**33.22**	64.80
GeoRA	**13.33**	9.17	73.40	**5.75**	**5.38**	33.04	**65.23**

across model scales. For reference, we also include the training cost of Qwen3-8B under our RLVR setting. In practice, randomized SVD reduces initialization time by roughly two orders of magnitude compared with standard SVD, making initialization a sub-minute operation even at 72B scale. Moreover, this memory cost is a one-off preprocessing overhead and can be released before training.

C Implementation Details

C.1 General Training Setup

We adopt GRPO as the default RLVR optimization algorithm for all experiments. To ensure fair comparison, we keep the training data, rollout configuration, and optimization budget identical across compared methods whenever applicable. Specifically, all PEFT methods use rank $r = 16$ and scaling factor $\alpha = 32$, with `target_modules` set to `all-linear`, i.e., adapters are applied to all linear layers. For GeoRA, we use sparsity ratio $\rho = 0.2$ unless otherwise specified. We train all models with AdamW in bfloat16 precision on 80GB GPUs. Unless otherwise specified, the learning rate is set to 1×10^{-6}, the global batch size is 128, and the rollout number is 8. When KL regularization is enabled, we use a shared reference policy and set the KL coefficient to 0.001. Unless otherwise specified, all methods are trained under the same decoding configuration and differ only in the adaptation parameterization based on a single random run.

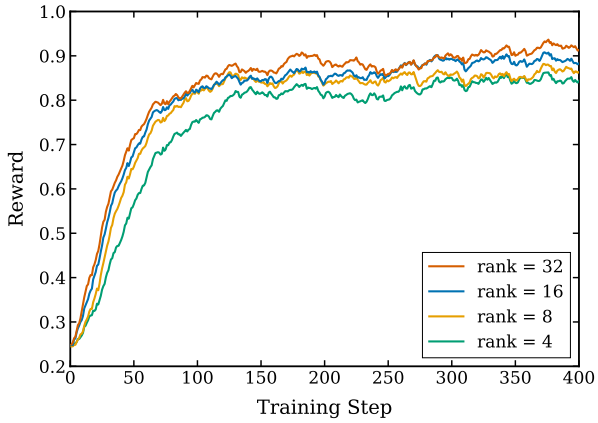
C.2 Task-Specific Details

Mathematical RLVR. For the main mathematical RLVR experiments, we fine-tune Qwen3-8B-Base and Llama-3.1-8B-Instruct on DeepMath-103K using GRPO. In addition, we evaluate smaller-scale settings on Qwen3-4B-Base and Qwen2.5-1.5B-Instruct using GSM8K to verify whether the gains of GeoRA persist across model scales. The maximum prompt length is set to 1024 tokens and the maximum response length is set to 4096 tokens. We train on 1 node with 8 GPUs. For DeepMath-103K, we use `math-verify` for answer verification by comparing the model output against the boxed ground-truth answer under expression- and LaTeX-based extraction rules. For GSM8K, we extract the final numerical answer from the model output and compare it against the ground-truth answer using exact match. Evaluation is conducted on AIME24, AIME25, MATH500, and OlymMATH.

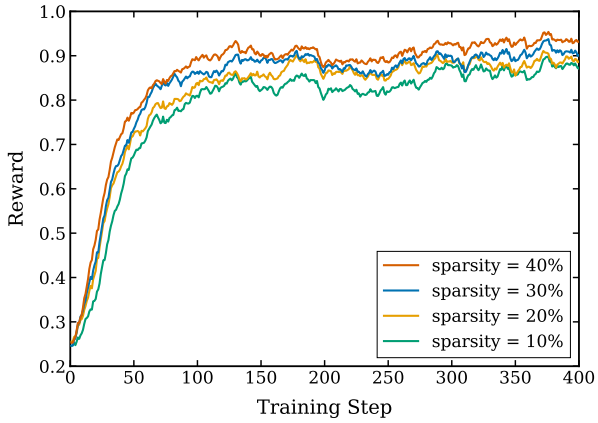
Medical RLVR. For medical reasoning, we train Llama-3.1-8B-Instruct on AlphaMed. We use the AlphaMed training split for RL training and the corresponding test split for validation. The maximum prompt length is set to 1024 tokens and the maximum response length is set to 4096 tokens. We train on 1 node with 8 GPUs. For reward computation, we extract the final predicted option from the model output, primarily using the `\boxed{X}` format and several common variants, and compare it against the ground-truth choice. A binary reward is assigned, with 1.0 for a correct answer and 0.0 otherwise. We report results on MedQA, MedM-

CQA, and PubMedQA.

Coding RLVR. For coding experiments, we train Qwen3-32B on Eurus-2-RL-Data. The maximum prompt length is set to 1024 tokens and the maximum response length is set to 2048 tokens. We train on 2 nodes with 8 GPUs per node. For reward computation, we extract the generated code from the model output and execute it against the provided test cases in a restricted execution environment. A sample is regarded as correct only when the generated program passes the corresponding test cases within the timeout limit. During evaluation, we follow the official benchmark protocol whenever available, including the same execution environment and decoding configuration across methods. We report results on LiveCodeBench, HumanEval, and MBPP.



(a) Sensitivity to Rank Variations



(b) Sensitivity to Sparsity Variations

Figure 7: Parameter Robustness Analysis.

C.3 Extended Experiments

We further evaluate GeoRA on additional backbones with different parameter scales (Qwen2.5-1.5B and Qwen3-4B) to assess scalability and generality. Table 8 reports both in-distribution reasoning performance and out-of-distribution generalization under the same evaluation protocol.

D Robustness Analysis

We evaluate the sensitivity of GeoRA's performance to variations in key hyperparameters on the Qwen3-4B model (Figure 7). (a) Reward trajectories under varying rank $r \in \{4, 8, 16, 32\}$ (with fixed sparsity $\rho = 20\%$). (b) Reward trajectories under varying sparsity levels $\rho \in \{10\%, 20\%, 30\%, 40\%\}$ (with fixed rank $r = 16$). In both scenarios, GeoRA demonstrates exceptional robustness, maintaining stable convergence and consistent high reward across a wide range of parameter settings, even at extremely low ranks or high sparsity levels.