FRoD: Full-Rank Efficient Fine-Tuning with Rotational Degrees for Fast Convergence

Guoan Wan[1], **Tianyu Chen**[1], **Fangzheng Feng**[2], **Haoyi Zhou**[1], **Runhua Xu**[1*]

[1]School of Computer Science and Engineering, Beihang University, China
[2]School of Electronic Information and Communications, Huazhong University of Science and Technoledgy, China
{gawan, tianyuc, haoyi, runhua}@buaa.edu.cn, fangzhengfeng@hust.edu.cn

Abstract

Parameter-efficient fine-tuning (PEFT) methods have emerged as a practical solution for adapting large foundation models to downstream tasks, reducing computational and memory costs by updating only a small subset of parameters. Among them, approaches like LoRA aim to strike a balance between efficiency and expressiveness, but often suffer from slow convergence and limited adaptation capacity due to their inherent low-rank constraints. This trade-off hampers the ability of PEFT methods to capture complex patterns needed for diverse tasks. To address these challenges, we propose FRoD, a novel fine-tuning method that combines hierarchical joint decomposition with rotational degrees of freedom. By extracting a globally shared basis across layers and injecting sparse, learnable perturbations into scaling factors for flexible full-rank updates, FRoD enhances expressiveness and efficiency, leading to faster and more robust convergence. On 20 benchmarks spanning vision, reasoning, and language understanding, FRoD matches full model fine-tuning in accuracy, while using only 1.72% of trainable parameters under identical training budgets.

Code — https://github.com/Bane-Elvin/AAAI2026_FRoD

Introduction

Large foundation models (Liu et al. 2019; Radford et al. 2021) have unlocked powerful representations across domains, but their ever-growing parameter scales bring drastic computational and memory costs for downstream adaptation. Parameter-efficient fine-tuning (PEFT) paradigms—spanning additive modules (Hu et al. 2023), selective tuning (Guo, Rush, and Kim 2021), and reparameterization (Aghajanyan, Zettlemoyer, and Gupta 2020)—mitigate this overhead by updating only sparse subsets of weights. Among these, LoRA (Hu et al. 2021) injects low-rank adapters to enable efficient fine-tuning at virtually no extra inference cost. Despite these advances, achieving faster convergence with fewer parameters remains an ongoing research challenge.

Prior approaches (Figure 1) like SVD-based LoRA (Sun et al. 2024; Lingam et al. 2024) demands prohibitive memory, forcing most variants (Meng, Wang, and Zhang 2024; Wang et al. 2025; Fan et al. 2025) to keep only the top rank-r directions—thereby restricting update rank and adaptation capacity. Shared random-basis PEFT methods (Kopiczko, Blankevoort, and Asano 2023; Koohpayegani et al. 2023) further cut parameters by reusing fixed random subspaces across layers, but all updates are confined to theses subspaces, limiting task-specific expressiveness and ultimately hampering final performance (Li, Han, and Ji 2024; Albert et al. 2025).

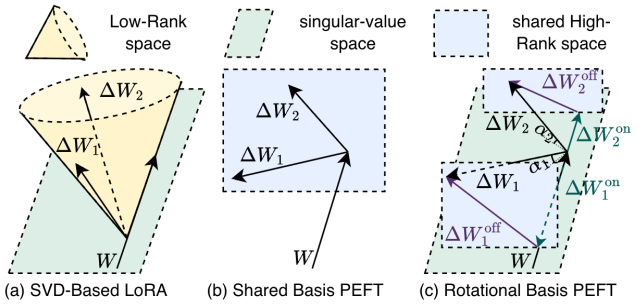


Figure 1: Update Space and Rank: FRoD (joint matrix decomposition and off-axis sparse spaces) vs. SVD-based and random PEFT methods.

We propose ***FRoD***, a full-rank, efficient fine-tuning method that couples strong expressiveness with fast convergence. FRoD starts by performing a hierarchical joint decomposition to extract a global shared basis and per-layer diagonal strength matrices Σ_i; updating only Σ_i delivers near-zero-cost on-axis fine-tuning. To step beyond those axes, we append a sparse, learnable matrix S_i to each layer. Its off-diagonal entries introduce off-axis interactions, injecting rotational degrees of freedom into the latent space (Figure 1c). This design maintains the foundational weight structure while offering flexible adaptation directions, supporting fast and stable optimization across diverse tasks.

- We propose a hierarchical joint decomposition method that extracts a shared latent basis across all model layer, capturing model-wide structure from the outset.
- We introduce sparse, learnable updates to enable flexi-

*Corresponding author

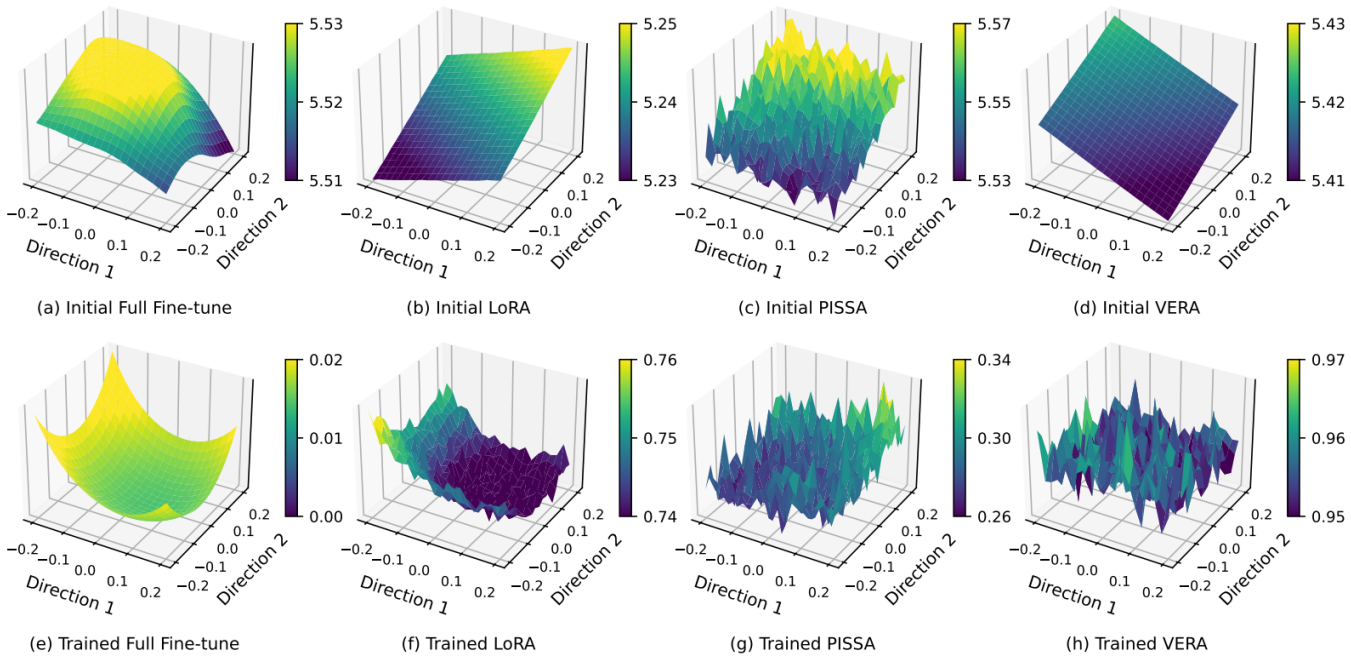


Figure 2: Comparison of loss landscapes for four representative fine-tuning methods in a principal parameter subspace. Each subplot shows the loss surface as a function of two principal directions (α and β) in parameter space, with the vertical axis denoting the loss value. The top row (a–d) presents landscapes at model initialization; the bottom row (e–h) shows them after convergence. These visualizations provide a comparative geometric view of how various fine-tuning strategies shape the optimization landscape during training.

ble off-axis adaptation, expanding the update space and improving both convergence and expressiveness.

- Extensive experiments on 20 vision, reasoning, and language benchmarks show that FRoD matches the accuracy of full fine-tuning, using just 1.72% of the parameters.

Background and Motivation

Background: LoRA and Its Variants

PEFT methods, such as LoRA and its variants, adopt a unified adaptation scheme expressed as $W = W_0 + \Delta W$, where W_0 denotes frozen pretrained weights and ΔW defines the learnable adaptation. Typically:

- *LoRA* replaces ΔW with the product of two trainable low-rank matrices, BA, reducing parameter costs.
- *VeRA* extends LoRA by introducing learnable scaling vectors with shared random matrices, $\Delta W = \Lambda_b B \Lambda_d A$, further decreasing redundancy in optimizer states.
- *PiSSA* imposes structural priors via truncated SVD: $W_0 = USV^\top$, initializing $B = U_{[:,:r]} S^{1/2}_{[:r,:r]}$ and $A = S^{1/2}_{[:r,:r]} V^\top_{[:,:r]}$ using dominant singular directions to facilitate convergence along principal axes.

Figure 2 visualizes the loss landscapes under different PEFT schemes, highlighting their geometric properties before and after fine-tuning. This illustration motivates key aspects of our method; owing to space limitations, detailed discussion of related work is relegated to the Appendix A & B.

Motivation: Convergence Efficiency

The convergence efficiency of PEFT schemes is tightly linked to their loss landscape geometry at initialization. As shown in Figure 2b, LoRA's low-norm random initialization yields a smooth, isotropic loss surface, supporting larger learning rates and stable early optimization. In contrast, PiSSA (Figure 2c) aligns updates with dominant singular directions, sharpening the curvature and constraining the attainable step size, often resulting in slower convergence. Despite distinct initializations, both LoRA and PiSSA ultimately converge to similarly sharp optima (Figure 2f–g), reflecting the inherent limitations of low-rank update spaces for capturing task complexity. In contrast, VeRA (Figure 2d) sustains a smooth loss surface akin to LoRA, enabling aggressive learning rates.

These observations emphasize the critical role of structured initialization in shaping early convergence dynamics. Inspired by these findings, we propose a hierarchical joint decomposition framework that globally extracts shared subspaces. This approach supports invariant, structure-aware initialization across layers, providing a principled trade-off between convergence speed and parameter efficiency.

Motivation: Parameter Degree of Freedom

Full fine-tuning enables unrestricted optimization in the entire parameter space, exhibiting superior convergence speed and expressive capacity. In contrast, PEFT methods, despite their lower training and deployment costs, are often confined to limited update subspaces. This constraint not only

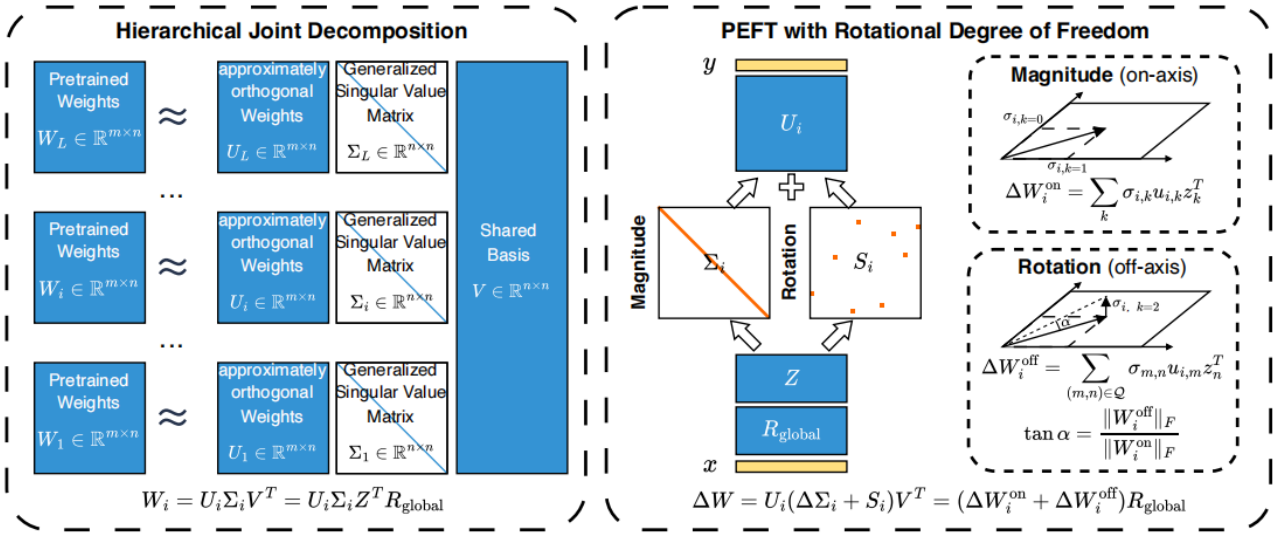


Figure 3: Overview of FRoD. FRoD is a two-stage approach: hierarchical joint decomposition initializes the model by extracting a global shared basis, and sparse perturbations introduce rotational degrees of freedom during fine-tuning.

impairs the linear descent trajectory but also compresses the representational flexibility in later stages, hindering adaptation to complex tasks. As depicted in Figure 2, LoRA and VeRA share similar (zero) initialization but differ in update space geometry: LoRA preserves relative smoothness and flexibility (Figure 2f), whereas VeRA's highly constrained subspace results in a distorted loss landscape and reduced convergence stability (Figure 2d).

These phenomena suggest that it is not initialization alone but, more fundamentally, the dimensionality of the effective update space, often termed the Parameter Degrees of Freedom (PDoF), that determines adaptation capability. Enhancing PDoF requires expanding the space of feasible updates, preferably in a principled, structured way. We operationalize this via rotations in both the global shared basis and its orthogonal complements. While direct rotation via dense orthogonal matrices is impractical in high dimensions, we approximate this mechanism through additive sparse matrices, which efficiently inject rotational flexibility with minimal computational overhead.

Method

FRoD enhances parameter-efficient fine-tuning through a two-stage framework designed to maximize both convergence speed and expressiveness, as illustrated in Figure 3. First, we perform a hierarchical joint decomposition that uncovers a globally shared subspace across layers, while preserving complementary layer-specific orthogonal components, yielding a principled and unified parameter initialization. Building upon this foundation, FRoD introduces a novel rotational PEFT mechanism that decomposes parameter updates into magnitude changes along the original directions and sparse off-axis rotations. This design injects rotational degrees of freedom into the update space, effectively expanding model expressiveness while maintaining parameter efficiency.

Hierarchical Joint Decomposition

We propose a hierarchical joint decomposition-based initialization scheme by performing joint decomposition (Kempf, Goulart, and Duncan 2023) across layers and categories.

Given a set of pre-trained matrices $\{W_i^{(c)}\}$, where i indexes the layer and $c \in \mathbb{C}$ denotes the category, our objective is to extract a shared latent basis for subsequent fine-tuning.

For each category c, we concatenate the weight matrices $W_i^{(c)} \in \mathbb{R}^{m\times n}$ from all L layers vertically:

$$W^{(c)} = \left[W_1^{(c)} W_2^{(c)} \cdots W_L^{(c)}\right]^T \in \mathbb{R}^{M_c \times n}, \tag{1}$$

where $M_c = L \cdot m$.

A thin QR decomposition yields:

$$W^{(c)} = Q^{(c)} R_{\text{global}}, \tag{2}$$

with $Q^{(c)} \in \mathbb{R}^{M_c \times n}$ orthonormal and $R_{\text{global}} \in \mathbb{R}^{n\times n}$ upper triangular. Each $Q^{(c)}$ is partitioned into layer-wise blocks Q_i according to the row dimensions of $W_i^{(c)}$.

Next, we regularize and aggregate the category-wise projections by computing:

$$T_\pi = \frac{1}{|\mathbb{C}|} \sum_{c\in\mathbb{C}} \left((Q^{(c)})^\top Q^{(c)} + \pi I\right)^{-1}, \tag{3}$$

where $\pi > 0$ is a small constant to stabilize the inversion. We diagonalize T_π as:

$$T_\pi = Z \Lambda Z^\top, \tag{4}$$

where $Z \in \mathbb{R}^{n\times n}$ is orthogonal. The resulting shared latent basis is given by:

$$V^\top = Z^\top R_{\text{global}}. \tag{5}$$

For each layer-category pair, we compute the local full-rank representation:

$$B_i = Q_i Z, \quad \Sigma_i = \text{diag}(\|B_i\|_2), \quad U_i = B_i \Sigma_i^{-1}, \tag{6}$$

	FRoD	LoRA	VeRA
Weights	$L(mn + s\,n^2 + n) + n^2$	$L(mn + mr + nr)$	$L(mn + r + n) + mr + nr$
Optim States	$2L(s\,n^2 + n)$	$2L(mr + nr)$	$2L(r + n)$

Table 1: Trainable parameter and optimizer states comparison. Assume $W \in \mathbb{R}^{m\times n}$, rank r.

which yields the approximation:

$$W_i^{(c)} \approx U_i \Sigma_i V^\top. \tag{7}$$

This hierarchical joint decomposition extracts a global orthonormal basis V across categories and layers, while the local orthogonal components U_i and scaling factors Σ_i retain layer-specific information. The reconstructed factors provide high-fidelity initialization for our fine-tuning modules.

PEFT with Rotational Degree of Freedom

Building upon the joint decomposition described in the previous section, we propose a PEFT method that augments the factorized form with a learnable sparse perturbation. Instead of directly using the decomposition in Eq. 7, We introduce a new trainable matrix $S_i \in \mathbb{R}^{n\times n}$ and redefine the adaptation matrix as

$$W_i' = U_i(\Sigma_i + S_i)V^\top, \tag{8}$$

where S_i adopts random off-diagonal sparsity matrix, and $s = \mathrm{nnz}(S_i)/n^2$ denotes its density, enabling rotation in the latent space with minimal parameter overhead.

Rotational Degree of Freedom(RDoF). The sparse perturbation S_i introduces a RDoF in the latent space. For an input vector $x \in \mathbb{R}^n$, the original transformation is:

$$y = W_i x = U_i \Sigma_i V_i^\top x. \tag{9}$$

Let $Z \in \mathbb{R}^{n\times n}$ be a transformation matrix, and define the latent coordinates as $c_v = R_{\text{global}} x$, where $R_{\text{global}} \in \mathbb{R}^{n\times n}$ is an orthogonal basis. The transformation becomes:

$$y = U_i \Sigma_i Z^\top c_v, \tag{10}$$

representing an axis-aligned scaling in the latent space.

With perturbation S_i, the adapted transformation becomes

$$y' = W_i' x = U_i(\Sigma_i + S_i)V_i^\top x = U_i(\Sigma_i + S_i)Z^\top c_v. \tag{11}$$

Thus, the update matrix is

$$\Delta W_i = W_i' - W_i = U_i(\Delta\Sigma_i + S_i)Z^\top R_{\text{global}}, \tag{12}$$

which can be decomposed into orthogonal components:

- **Magnitude (on-axis)**: $\Delta W_i^{\text{on}} = U_i \Delta\Sigma_i Z^\top$, adjusting the generalized singular values (intensity).
- **Rotation (off-axis)**: $\Delta W_i^{\text{off}} = U_i S_i Z^\top$, introducing cross-dimensional coupling via the off-diagonal S_i.

These components satisfy an orthogonality condition and thereby enable a geometric decomposition of the update. By explicitly modeling this rotational degree of freedom with a sparse, learnable matrix, our method significantly increases the expressiveness of adaptation while retaining strong parameter efficiency.

Preservation of Singular-Value Spectrum. To provide a theoretical foundation for our rotational PEFT framework, we first establish a spectral-stability theorem for sparse perturbations, followed by two corollaries: (i) an orthogonal decomposition of on-axis and off-axis updates, and (ii) an angular representation of the total update. Collectively, these results show that sparse off-axis rotations keep the dominant singular-value directions essentially intact while expanding the update space in a controlled way.

Theorem 1 (Spectral Stability Under Sparse Perturbations)**.** *Let $\Sigma_i \in \mathbb{R}^{n\times n}$ be a diagonal matrix of singular values and let $S_i \in \mathbb{R}^{n\times n}$ be a perturbation such that*

$$\|S_i\|_0 \ll n^2 \quad \textit{and} \quad \max_{p,q} |s_{pq}| \le \varepsilon. \tag{13}$$

Then, for every k,

$$\big|\sigma_k(\Sigma_i + S_i) - \sigma_k(\Sigma_i)\big| \;\le\; \|S_i\|_2, \tag{14}$$

so the dominant singular directions encoded by Σ_i are preserved up to a deviation bounded by $\|S_i\|_2$.

Corollary 2 (Orthogonal Decomposition of Updates)**.** *Define the on-axis and off-axis updates as $\Delta W_i^{on},\ \Delta W_i^{off} \in \mathbb{R}^{m\times n}$. Then*

$$\langle \Delta W_i^{on}, \Delta W_i^{off} \rangle_F \;=\; 0, \tag{15}$$

i.e. the two update components are orthogonal in the Frobenius inner product.

Corollary 3 (Angular Representation of the Total Update)**.** *Let $\tan\alpha = \|\Delta W_i^{off}\|_F / \|\Delta W_i^{on}\|_F$ and set $\hat{U}_{on} = \Delta W_i^{on}/\|\Delta W_i^{on}\|_F,\ \hat{U}_{off} = \Delta W_i^{off}/\|\Delta W_i^{off}\|_F$. Then*

$$\Delta W_i = \|\Delta W_i\|_F \big(\cos\alpha\, \hat{U}_{on} + \sin\alpha\, \hat{U}_{off}\big). \tag{16}$$

For small α (when $\|\Delta W_i^{on}\|_F \gg \|\Delta W_i^{off}\|_F$),

$$\Delta W_i \approx \|\Delta W_i^{on}\|_F \left(\hat{U}_{on} + \alpha \cdot \hat{U}_{off}\right) R_{global}. \tag{17}$$

Thus, the total update can be interpreted as a rotation from the axis-aligned direction by a small angle α, reflecting a controllable expansion as Shared Random-Basis PEFT.

Detailed proofs are provided in the Appendix C.

Memory Usage of Projection Matrices

In FRoD, the factorized form Eq. 8 uses frozen matrices $U_i \in \mathbb{R}^{m\times n}$ and shared $V \in \mathbb{R}^{n\times n}$, while only $\Sigma_i \in \mathbb{R}^n$ and sparse $S_i \in \mathbb{R}^{n\times n}$ are trainable. Across L layers, the total number of trainable parameters is $L(mn+3sn^2+n)+n^2$. For comparison, LoRA with rank r has $L(mn + 3mr + 3nr)$ and VeRA shares fixed random bases $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$ across layers and trains only scaling vectors $d \in \mathbb{R}^r, b \in \mathbb{R}^n$, yielding $L(mn + 3r + 3n) + mr + nr$. A comparison between FRoD, LoRA and VeRA is shown in Table 1.

Method	# Params (%)	Cars	DTD	EuroSAT	GTSRB	RESISC45	SUN397	SVHN	Average
Full FT	100	60.33	73.88	98.96	98.30	93.65	53.84	96.78	82.25
LoRA †	1.49	41.02	70.15	98.66	96.51	90.38	47.51	95.39	77.09
PiSSA †	1.49	40.41	69.62	98.48	95.84	90.58	47.21	95.84	76.85
MiLoRA †	1.49	39.77	70.48	98.19	97.52	89.92	45.38	95.49	76.68
GOAT †	2.24	53.50	**75.32**	98.82	**98.17**	**93.46**	54.53	**96.62**	81.49
VeRA	0.29	60.37	73.03	98.44	97.30	89.71	50.25	95.93	80.71
RandLoRA	1.49	42.36	69.68	98.37	96.88	89.03	47.47	95.60	77.06
DoRA †	1.49	40.75	71.91	**98.89**	97.71	90.19	47.54	95.46	77.49
FRoD(s=0.01)	1.29	61.30	71.22	98.52	97.66	91.43	**59.62**	96.01	82.25
FRoD(s=0.02)	2.49	**62.13**	73.24	**98.89**	97.84	91.95	59.16	96.42	**82.80**

Table 2: We evaluate CLIP ViT-B/32 across StanfordCars, DTD, EuroSAT, GTSRB, RESISC45, SUN397, and SVHN datasets. The symbol † indicates that the results are taken from (Fan et al. 2025).

Method	# Params(%)	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	OBQA	Average
ChatGPT †	/	73.10	85.40	68.50	78.50	66.10	74.80	77.01
Full FT	100	88.50	83.79	81.16	92.50	82.95	78.60	84.58
LoRA†	0.84	69.80	79.90	70.50	83.60	82.60	81.00	77.61
PiSSA†	0.84	67.60	78.00	71.80	78.00	75.80	75.60	73.78
MiLoRA†	0.84	67.60	83.30	73.40	88.20	83.00	80.80	79.24
GOAT †	0.96	73.60	**83.95**	80.55	80.50	**85.00**	**87.00**	82.73
VeRA	0.02	77.68	69.58	84.73	87.99	50.00	30.20	66.70
RandLoRA	0.84	87.65	82.70	88.05	92.91	75.77	34.20	76.88
DoRA†	0.84	71.80	83.10	79.90	89.10	83.00	81.20	80.45
FRoD(s=0.02)	0.18	**87.74**	83.62	**81.17**	**93.28**	80.98	74.40	**83.53**

Table 3: Performance comparison of LLaMA2 7B on 6 commonsense reasoning datasets. The symbol † indicates that the results are taken from (Fan et al. 2025).

Experiments

Baselines

We compare FRoD with a wide range of strong baselines, grouped into four categories, to comprehensively evaluate its effectiveness and robustness. **Basic Mehtods**: *Full FT* - fine-tunes all parameters; *LoRA* (Hu et al. 2021); *ChatGPT* (Brown et al. 2020). **SVD-Based LoRA Methods**: *PiSSA* (Meng, Wang, and Zhang 2024); *MiLoRA* (Wang et al. 2025); *GOAT* (Fan et al. 2025). **Shared Random-Basis PEFT Methods**: *VeRA* (Kopiczko, Blankevoort, and Asano 2023); *RandLoRA* (Albert et al. 2025). **Architecture-based Methods**: *DoRA* (Liu et al. 2024).

Datasets

We evaluate FRoD across 20 tasks, spanning 3 domains. **Image Classification (IC)**: We fine-tune and evaluate Clip ViT-B/32 (Radford et al. 2021) on 7 image classification datasets. **Commonsense Reasoning (CR)**: We fine-tune LLaMA2-7B (Touvron et al. 2023) on Commonsense170K and evaluate on 6 commonsense reasoning datasets (Hu et al. 2023). **Natural Language Understanding (NLU)**: We fine-tune RoBERTa-large (Liu et al. 2019) on 7 GLUE tasks (Wang et al. 2019), following the setup of (Hu et al. 2021).

Main Results

Tables 2, 3 and 4 present results on 3 domain benchmarks:

- **Image classification (Table 2).** FRoD attains the best *average* accuracy across seven datasets in just three epochs, edging out full fine-tuning by 0.4 accuacy with only 2.5% trainable parameters. GOAT outperforms FRoD on two individual datasets, a gap we attribute to its dataset-specific rank tuning for a detailed comparison.
- **Commonsense reasoning (Table 3).** Within four epochs, FRoD delivers decisive gains over full FT on BoolQ, PIQA, SIQA and HellaSwag while maintaining a 500× parameter reduction. Lower scores on WinoGrande and OBQA stem from limited training examples and prompt-template designing.
- **Natural language understanding (Table 4).** Running for only three epochs, FRoD matches or surpasses full FT on five GLUE tasks. The single shortfall on RTE is largely due to the tight epoch budget, extending training to ten epochs recovers parity.

In summary, FRoD consistently delivers leading performance across all benchmarks, excels in nearly every subtask, and narrows the remaining gap to full fine-tuning and,

Method	# Params (%)	CoLA	SST-2	MRPC	QQP	MNLI	QNLI	RTE	Average
Full FT	100	84.27	96.44	90.44	91.58	90.84	94.58	84.84	90.42
LoRA †	4.00	83.41	95.64	83.33	90.06	89.00	93.28	84.47	88.46
PiSSA †	4.00	69.12	95.98	82.84	91.24	88.94	93.59	73.29	85.00
MiLoRA †	4.00	84.65	96.10	86.02	91.33	89.51	94.12	84.83	89.51
GOAT †	4.50	**86.86**	96.21	84.55	91.40	89.55	94.19	**85.56**	89.76
VeRA	0.22	81.78	95.76	86.59	89.30	89.48	94.39	75.09	87.48
RandLoRA	0.84	84.66	96.10	89.71	90.24	**89.58**	94.25	80.14	89.24
DoRA †	4.00	85.33	95.99	84.07	91.24	89.52	93.54	84.48	89.17
FRoD(s=0.02)	2.49	**86.86**	**96.33**	**90.44**	**91.63**	**89.58**	**94.40**	84.48	**90.53**

Table 4: Performance comparison of RoBERTa-large with different methods on 7 GLUE tasks. The symbol † indicates that the results are taken from (Fan et al. 2025).

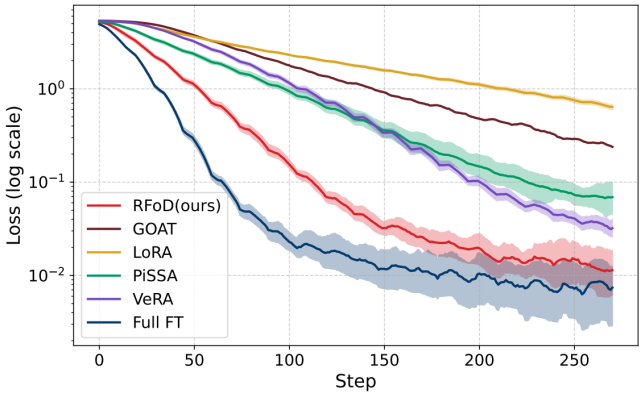


Figure 4: Training loss curves of Different LoRA methods and Full Fine-tuning on Cars. The shaded areas in the figure represent the error bounds of different methods. We randomly sample five seed values.

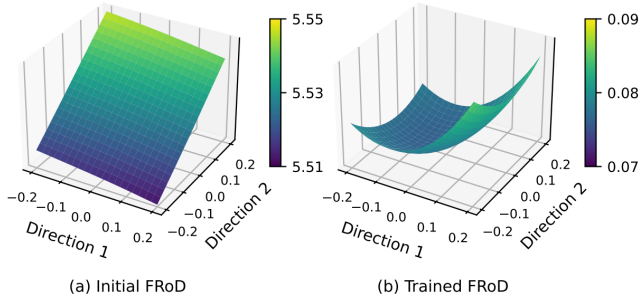


Figure 5: Visualization of Loss Landscapes for FRoD in Model Parameter Space.

in some cases, even eliminates it, thereby attesting to the heightened efficacy of our method. All metrics are averaged over three independent training runs on each dataset. Further prompt-template details and analyses are presented in Appendix D.

Convergence Speed

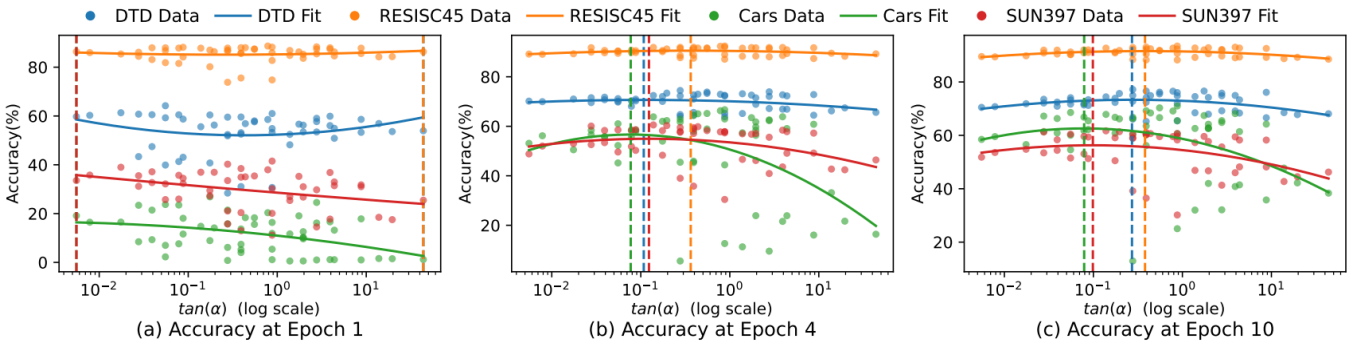
FRoD exhibits markedly faster convergence across the main experiments. It reaches convergence within 4 epochs on most IC benchmarks and within a single epoch on SVHN, whereas the strongest LoRA-based baseline, GOAT, requires more than ten epochs. On CR and NLU tasks, FRoD achieves peak validation accuracy in 3 epochs, except for RTE, which stabilizes after ten, whereas competing methods typically require between three and one hundred epochs. The complete set of training hyper-parameters is provided in the Appendix D.

Figure 4 illustrates the epoch-wise validation accuracy on Stanford Cars for FRoD under five different random seeds. The solid curve denotes the mean accuracy across seeds, while the shaded region represents the min–max range of those runs. Across these runs, FRoD's final accuracy remains within 0.1% of full fine-tuning, demonstrating that our convergence speed and performance are stable and insensitive to initialization variance.

Analysis of Loss Landscape Geometry

To elucidate why FRoD achieves superior convergence, we visualize its loss landscape in Figure 5. As quantified in the Appendix D.5, our hierarchical joint decomposition reconstructs each layer with a maximum error below 1e-5, effectively preserving the complete generalized singular-value spectrum. This initialization results in a smooth and isotropic initial loss surface (Figure 5a). In contrast, PiSSA's reliance on dominant components leads to a more fragmented initial geometry (Figure 2c), while VeRA exhibits higher initial curvature (Figure 2d).

After optimization, FRoD (Figure 5b) maintains a well-conditioned, convex landscape that closely resembles full fine-tuning (Figure 2e), despite using significantly fewer trainable parameters. This stability confirms that the rotational degrees of freedom in FRoD provide sufficient flexibility for full-rank updates without inducing the pronounced curvature distortion seen in VeRA (Figure 2h). Ultimately, these geometric advantages allow FRoD to support aggressive learning rates and achieve faster, more robust convergence across diverse tasks. Further experiments on additional initialization methods are provided in Appendix D.6.



(a) Accuracy at Epoch 1 (b) Accuracy at Epoch 4 (c) Accuracy at Epoch 10

Figure 6: Impact of off-axis rotation angle α on classification accuracy. Scatter plots and fitted trends illustrate how varying the rotation angle and its magnitude affect accuracy at epochs 1, 4, and 10 on four vision datasets.

Hyperparameters			Avg. Acc.(Epoch) (%)		
s	lr(S)	lr(Σ_i)	1	4	10
0.0	0	1e-04	25.89	55.89	70.77*
0.0	0	5e-04	34.20	66.52	**71.26**
0.0	0	1e-03	42.13	**68.58**	70.86
0.0	0	5e-03	**50.03**	65.66	67.28
0.01	1e-04	0	47.72	**71.33**	72.88
0.02	1e-05	0	29.96	62.47	70.77
0.02	5e-05	0	44.51	68.87	70.92
0.02	1e-04	0	**49.62**	70.89	**73.12**
0.02	5e-04	0	41.59	55.28	60.60
0.1	1e-04	0	46.95	70.68	72.77
0.1	5e-04	5e-03	37.47	48.51	52.22
0.02	1e-04	1e-03	48.36	**71.85**	**74.59**

Table 5: Hyperparameter sensitivity of FRoD on four vision datasets. The table reports average accuracy (%) at epochs 1, 4, and 10 for combinations of sparsity $s \in \{0.01, 0.02, 0.1\}$, off-axis learning rate $lr(S) \in \{$1e-5, 5e-5, 1e-4, 5e-4$\}$, and on-axis learning rate $lr(\Sigma_i) \in \{$1e-4, 5e-4, 1e-3, 5e-3$\}$. Rows are shaded to indicate the best (pink), second best (light pink), and worst (light gray) hyperparameter settings. An asterisk (*) denotes configurations that fail to converge on most datasets within 10 epochs.

Ablation Study

To better understand the roles of on-axis and off-axis updates, we conduct a detailed ablation study, separately disabling each component in FRoD.

Impact of On-axis. When removing the off-axis updates, we observe that increasing the learning rate for the dominant singular directions (Σ) significantly accelerates early-stage convergence—as reflected in high accuracy at epoch 1, as Table 5. However, overly large values introduce instability and lead to suboptimal final performance by epoch 10. This reveals a clear trade-off between convergence speed and stability for the on-axis learning rate.

Impact of Off-axis. In contrast, when removing the on-axis component, the off-axis updates alone offer higher training stability due to their broader parameter search space. Yet, large off-axis learning rates do not guarantee better performance, as these directions are not strictly orthogonal and can easily distort the update. Consequently, smaller learning rates are preferred for off-axis updates to preserve stability.

Impact of learning rate. Importantly, none of the ablated configurations outperforms the jointly activated setting, which combines both components and achieves the highest accuracy. Nonetheless, the overall performance remains sensitive to different learning rate combinations. Some poorly tuned configurations even underperform the ablated variants.

To investigate this further, we approximate, using learning rate, the rotation angle α between the on-axis and off-axis components, defined by Eq. 14. As shown in Figure 6, Early-stage performance (epoch 1) benefits from more aggressive updates in either axis. However, the best final accuracy (epoch 10) emerges when $\tan\alpha \in [0.05, 0.2]$, validating our theoretical insight that moderate rotation preserves strength alignment as Eq.17. Appendix E provides implementation details. This analysis confirms that well-calibrated rotational degrees of freedom not only enhance early convergence but also maintain representation stability throughout training.

Conclusion

We present FRoD, a novel PEFT approach that synergistically integrates hierarchical orthogonal decomposition and sparse perturbations to achieve both rapid convergence and high expressiveness. By extracting shared latent bases across layers and injecting rotational degrees of freedom via sparse matrices, FRoD substantially expands the effective update space at minimal parameter cost. Empirical results on 20 tasks across vision, reasoning, and language understanding demonstrate that FRoD consistently achieves or surpasses full fine-tuning and leading PEFT baselines, converging within only 1–4 epochs and exhibiting strong robustness to random initialization. These findings highlight FRoD's capability to reconcile parameter efficiency with adaptable optimization dynamics. As future work, we will explore leveraging the high-rank sparse subspaces induced by FRoD for continual learning and model merging, potentially enabling more efficient knowledge integration and mitigating catastrophic forgetting.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62302022, 62202029), and Young Elite Scientists Sponsorship Program by CAST (NO.20230NRC001).

References

Aghajanyan, A.; Zettlemoyer, L.; and Gupta, S. 2020. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. ArXiv:2012.13255.

Albert, P.; Zhang, F. Z.; Saratchandran, H.; Rodriguez-Opazo, C.; Hengel, A. v. d.; and Abbasnejad, E. 2025. RandLoRA: Full-rank parameter-efficient fine-tuning of large models. ArXiv:2502.00987.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Fan, C.; Lu, Z.; Liu, S.; Qu, X.; Wei, W.; Gu, C.; and Cheng, Y. 2025. Make LoRA Great Again: Boosting LoRA with Adaptive Singular Values and Mixture-of-Experts Optimization Alignment. ArXiv:2502.16894.

Guo, D.; Rush, A.; and Kim, Y. 2021. Parameter-Efficient Transfer Learning with Diff Pruning. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4884–4896. Online: Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. ArXiv:2106.09685.

Hu, Z.; Wang, L.; Lan, Y.; Xu, W.; Lim, E.-P.; Bing, L.; Xu, X.; Poria, S.; and Lee, R. K.-W. 2023. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. ArXiv:2304.01933.

Kempf, I.; Goulart, P. J.; and Duncan, S. R. 2023. A Higher-Order Generalized Singular Value Decomposition for Rank-Deficient Matrices. *SIAM Journal on Matrix Analysis and Applications*, 44(3): 1047–1072.

Koohpayegani, S. A.; Navaneet, K.; Nooralinejad, P.; Kolouri, S.; and Pirsiavash, H. 2023. NOLA: Compressing LoRA using Linear Combination of Random Basis.

Kopiczko, D. J.; Blankevoort, T.; and Asano, Y. M. 2023. VeRA: vector-based random matrix adaptation.

Li, Y.; Han, S.; and Ji, S. 2024. VB-LoRA: Extreme Parameter Efficient Fine-Tuning with Vector Banks.

Lingam, V.; Tejaswi, A.; Vavre, A.; Shetty, A.; Gudur, G. K.; Ghosh, J.; Dimakis, A.; Choi, E.; Bojchevski, A.; and Sanghavi, S. 2024. SVFT: Parameter-Efficient Fine-Tuning with Singular Vectors. *Advances in Neural Information Processing Systems*, 37: 41425–41446.

Liu, S.-Y.; Wang, C.-Y.; Yin, H.; Molchanov, P.; Wang, Y.-C. F.; Cheng, K.-T.; and Chen, M.-H. 2024. DoRA: Weight-Decomposed Low-Rank Adaptation.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692.

Meng, F.; Wang, Z.; and Zhang, M. 2024. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. ArXiv:2404.02948.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. PMLR. ISSN: 2640-3498.

Sun, C.; Wei, J.; Wu, Y.; Shi, Y.; He, S.; Ma, Z.; Xie, N.; and Yang, Y. 2024. SVFit: Parameter-Efficient Fine-Tuning of Large Pre-Trained Models Using Singular Values. ArXiv:2409.05926.

Sun, Y.; Chen, Q.; He, X.; Wang, J.; Feng, H.; Han, J.; Ding, E.; Cheng, J.; Li, Z.; and Wang, J. 2022. Singular Value Fine-tuning: Few-shot Segmentation requires Few-parameters Fine-tuning. *Advances in Neural Information Processing Systems*, 35: 37484–37496.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. ArXiv:1804.07461.

Wang, H.; Li, Y.; Wang, S.; Chen, G.; and Chen, Y. 2025. MiLoRA: Harnessing Minor Singular Components for Parameter-Efficient LLM Finetuning. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4823–4836. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.