SIMILAR FACTORS, DIFFERENT UPDATES: RETHINKING WHAT TO SHARE IN FEDERATED LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank adaptation (LoRA) enables efficient fine-tuning of foundation models in federated learning. However, determining what to share across clients remains challenging under data heterogeneity. Factor similarity provides an intuitive basis for selective sharing, but similar factors may reflect common initialization and accumulated server history rather than shared knowledge. We show that a single gradient step can produce perfectly aligned factors with opposite effective updates, revealing a fundamental limitation of similarity-based interpretations. Motivated by this finding, we propose GReP-LoRA, which uses task gradients to guide shared and personalized adaptation. Our method initializes a shared adapter from the population gradient and selects frozen private spectral bases from client residual gradients. Clients jointly learn shared tangent increments and compact private cores that rescale and combine the selected directions, while only shared parameters are aggregated. Our theoretical analysis connects captured residual energy to personalized objective improvement under quadratic assumptions and establishes nonconvex stationarity bounds with explicit optimization errors. We evaluate our method on vision and language tasks. Held-out gradient prediction and private-branch transfer experiments demonstrate client-specific specialization, while controlled comparisons support the usefulness of a full trainable core.

1 INTRODUCTION

Federated learning enables multiple clients to train a model while retaining their data locally (McMahan et al., 2017; Liang et al., 2025). Low-rank adaptation (LoRA) makes this approach more practical for foundation models by restricting trainable weight updates to a product of two small matrices, BA (Hu et al., 2022). Federated LoRA combines parameter-efficient adaptation with collaborative learning (Zhang et al., 2024), but heterogeneous client objectives introduce a further design question: *which adaptation directions should be shared, and which should remain private?*

An appealing intuition follows the computation BAx: A first projects the input into a low-dimensional space, and B maps the resulting representation back to the layer's output. Because B acts closer to the adapter output, it may seem better suited to task-specific adaptation. FedSA-LoRA associates A with general knowledge and B with client-specific knowledge, and accordingly shares A while retaining B locally (Guo et al., 2025). Its observed cross-client factor similarities support this interpretation in the studied setting. Yet factor position alone does not establish a universal division of learning roles, especially because LoRA's initialization treats the two factors differently (Hayou et al., 2024).

Our diagnostic in Figure 1 directly examines this ambiguity. Under standard FedIT initialization, B exhibits larger normalized same-client changes and lower early cross-client similarity than A. Reversing the zero/nonzero initialization reverses both patterns: A becomes the more variable and less aligned factor, while B remains highly similar across clients (Figure 1 (a,b,e,f)). The apparently more personalized side therefore depends on the initialization scheme. Moreover, the green curves show that both complete shared factors of our method remain highly aligned, even though their changes ΔA and ΔB exhibit substantially lower agreement in early rounds (Figure 1 (e-h)). Thus, similarity of complete factors can conceal differences in what clients are learning. Our theoretical counterexample (Proposition 11 in Appendix N.1) makes this distinction precise: one factor-SGD step from a common nonzero initialization on two convex quadratic client objectives can yield perfectly aligned factors but opposite effective weight increments. These observations motivate choosing

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

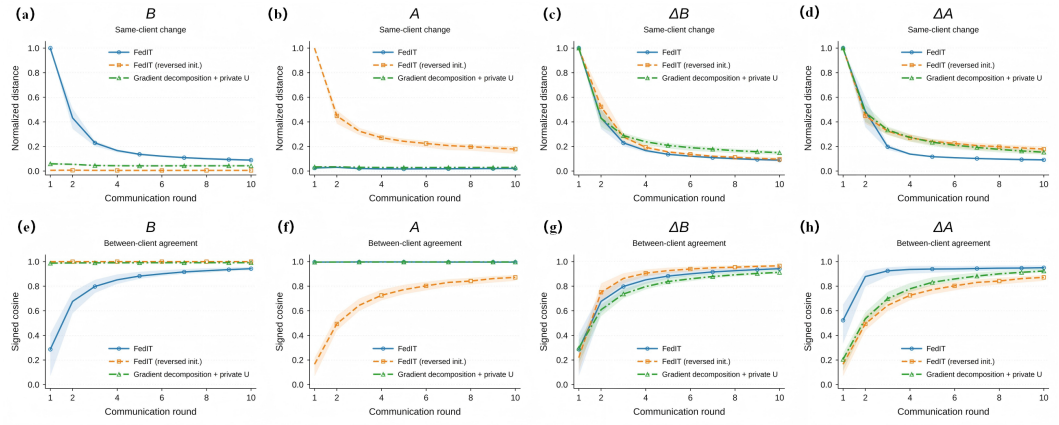


Figure 1: Initialization-dependent asymmetry in federated LoRA. Top: same-client normalized distances for B, A, ΔB, and ΔA. Bottom: cross-client signed cosine similarities. Reversing the initialization reverses the apparent asymmetry between A and B. Highly similar complete factors can coexist with substantially less aligned factor changes in early communication rounds.

shared and private adaptation directions from client objectives. We introduce *Gradient Residual Personalization for LoRA* (GReP-LoRA), which makes this choice at initialization. At a common pretrained model, let G_i be the gradient of client i's loss with respect to an adapted weight matrix, and let p_i be its normalized training-sample weight. We decompose

$$\bar{G} = \sum_i p_i G_i, \qquad R_i = G_i - \bar{G}. \tag{1}$$

The population gradient $\bar{G}$ describes the first-order behavior of the weighted global objective, whereas R_i measures the client's departure from that objective at the same model. This provides a task-dependent criterion for allocating adaptation capacity before federated training begins.

GReP-LoRA uses $\bar{G}$ to initialize a shared adapter through gradient-based initialization (Wang et al., 2024a). For each client, the leading singular directions of R_i define frozen private bases, and a small trainable core learns how to combine those directions using the local task loss. The core parameterization builds on fixed-basis adaptation (Bałazy et al., 2024; Ponkshe et al., 2024); our design selects its bases from client gradient residuals and retains its coefficients locally. Both shared factors are trained and aggregated, while private cores remain on their clients. Personalization can use both input and output directions within a space selected for each client's initial objective contrast.

The shared update also accounts for the bilinear structure of LoRA. Around common round anchors A and B, full factor updates contain both the first-order term $B\delta A_i + \delta B_i A$ and the second-order interaction $\delta B_i \delta A_i$. Averaging factors generally fails to preserve averaged products, an issue addressed by prior aggregation methods (Singhal et al., 2025b). GReP-LoRA omits the local cross term and jointly trains the resulting first-order shared update with the private core. Since clients use common anchors, averaging factor increments also averages their first-order weight-space effects. Client-specific adaptation is learned through the private branch. Server reconstruction of the shared factor product still introduces a second-order term, which we explicitly retain in our analysis.

Our analysis establishes the residual-approximation property of the selected spectral spaces and relates captured residual energy to attainable personalized objective improvement under stated quadratic curvature assumptions. We evaluate GReP-LoRA on DomainNet (Peng et al., 2019), NICO++ (Zhang et al., 2023), and GLUE (Wang et al., 2019), covering vision and language tasks under heterogeneous federated settings. Improvements over FedSA-LoRA in the reported NICO++ and GLUE comparisons, together with our initialization diagnostics, support objective-based direction selection for federated personalization. Our contributions are threefold:

- We show that initialization can reverse the apparent asymmetry between LoRA factors, and establish that factor agreement alone cannot certify agreement in effective updates.
- We propose GReP-LoRA, which combines population-gradient shared initialization, residual-gradient private spaces, and locally linear shared updates with trainable private cores.

- We connect residual-space selection to conditional objective improvements and evaluate the method across vision and language tasks, with diagnostics that distinguish complete-factor similarity from agreement in learned changes.

2 Related Work

2.1 Federated LoRA aggregation and selective sharing.

Federated LoRA methods address the interaction between factorized local training, server aggregation, and heterogeneous client objectives. FFA-LoRA freezes a common factor and trains only the other factor, simplifying aggregation and reducing communication (Sun et al., 2024). FLoRA uses stacking-based aggregation to preserve client adapter products, while FlexLoRA aggregates effective updates and redistributes them through rank-dependent SVD truncation (Wang et al., 2024b; Bai et al., 2024). FedEx-LoRA compensates for the discrepancy between factor averaging and product averaging, and LoRA-FAIR jointly addresses aggregation bias and initialization across communication rounds (Singhal et al., 2025b; Bian et al., 2025). Beyond aggregation, a related line of work considers which parameters should be shared for personalization. FedSA-LoRA associates the two factors with general and client-specific knowledge and aggregates only the shared A factor (Guo et al., 2025). FedAS-LoRA selects the sharing side before training using a rank-aware shared-subspace criterion (Xu et al., 2026), while FedDPA combines global and local adapters through instance-dependent weighting at inference (Yang et al., 2024).

Our work examines whether cross-client factor similarity provides sufficient evidence for sharing. We show that initialization and accumulated server history can produce factor agreement without agreement in current effective updates. Motivated by this observation, GReP-LoRA uses population and residual gradients to select shared and private adaptation spaces. Its focus is on determining which adaptation directions support shared or personalized learning, complementing prior work on aggregation fidelity and factor-level sharing.

2.2 Gradient-informed and geometric low-rank adaptation.

The behavior of low-rank adaptation depends on its initialization, update space, and optimization dynamics. Analyses of LoRA initialization show that reversing the zero and nonzero factors can lead to different learning dynamics despite an identical initial effective model (Hayou et al., 2024). PiSSA initializes adapters from the principal components of pretrained weights, whereas LoRA-GA uses task gradients to align the initial induced update with full fine-tuning (Meng et al., 2024; Wang et al., 2024a). Fixed-basis methods further reduce the trainable parameter count: LoRA-XS learns a small matrix between frozen spectral factors, and LoRA-SB combines this architecture with initialization based on approximating the initial gradient (Bałazy et al., 2024; Ponkshe et al., 2024). Fed-SB extends this parameterization to federated learning, using common frozen factors so that averaging client cores exactly averages the represented updates (Singhal et al., 2025a). Geometric approaches also study updates in the induced weight space. ISO-LoRA couples factor updates through joint tangent-space optimization to improve rank utilization, while RAFFT combines Riemannian parameter matching with rank-preserving local optimization in federated fine-tuning (Zhu et al., 2026; Zhou et al., 2026).

GReP-LoRA builds on these established ingredients to allocate adaptation directions between shared and personalized learning. It selects private bases from each client's residual relative to the population gradient, allowing the private adaptation space to reflect client-specific gradient information. The corresponding private cores remain local and learn both the amplitudes and combinations of the selected directions, alongside shared tangent increments. The contribution therefore lies in the allocation of shared and private adaptation spaces; gradient initialization, fixed-basis parameterizations, and tangent-space optimization provide the foundations for this formulation.

3 Method

We consider N clients, where client i has an empirical loss f_i and n_i training examples, with weight $p_i = n_i / \sum_j n_j$. GReP-LoRA allocates adaptation capacity to a shared low-rank component and a

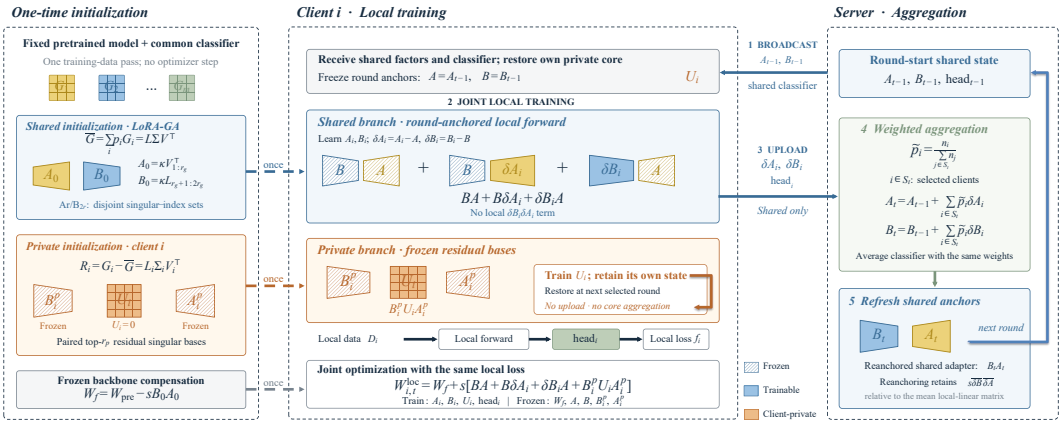


Figure 2: **Overview of GRePP-LoRA. Left:** One-time gradient calibration uses the weighted mean client gradient to initialize shared factors and client-specific residuals to select frozen private bases. Zero-initialized private cores and backbone compensation preserve the initial pretrained effective weight. **Center:** Clients jointly optimize a round-anchored first-order shared model, their private cores, and the classifier using local task losses. **Right:** The server aggregates shared factor increments and the classifier using sample-count weights, then updates the shared anchors for the next round. Private bases remain frozen, and private cores are retained locally without aggregation.

client-specific correction. The initialization diagnostic in Figure 1 motivates choosing this allocation from task gradients at a common model.

For one adapted layer, the personalized evaluation model at round t is

$$W_{i,t}^{\text{eval}} = W_f + s\left(B_t A_t + B_i^p U_{i,t} A_i^p\right). \tag{2}$$

Here, W_f is a compensated frozen backbone weight, $A_t \in \mathbb{R}^{r_g \times d_{\text{in}}}$ and $B_t \in \mathbb{R}^{d_{\text{out}} \times r_g}$ are shared trainable factors, and $A_i^p \in \mathbb{R}^{r_p \times d_{\text{in}}}$ and $B_i^p \in \mathbb{R}^{d_{\text{out}} \times r_p}$ are frozen private bases. Each client retains a trainable full core $U_{i,t} \in \mathbb{R}^{r_p \times r_p}$. We use $s = \alpha / r_g$ and apply this construction to every adapted layer. A shared classifier is trained and aggregated with the shared factors.

Overview. Figure 2 illustrates the workflow of GRePP-LoRA. During one-time calibration at a common pretrained model with a common classifier, the weighted mean client gradient initializes the shared factors, while each client's gradient residual determines its frozen private bases. The private cores start at zero, and compensating the frozen backbone preserves the pretrained effective weight in the initial evaluation model. In each communication round, participating clients receive the shared factors and classifier and restore their own private cores. With the received factors fixed as round anchors, each client jointly trains the shared factor increments, private core, and classifier using its local task loss. The shared branch uses a first-order expansion around these anchors, whereas the private branch adapts within the residual-selected bases. The server then averages the shared increments and classifier with sample-count weights and updates the shared factors for the next round. Private bases remain frozen, and the updated cores are retained locally for future participation. We next detail the initialization and round-wise optimization, followed by an analysis of the private-space selection criterion.

3.1 Gradient-Guided Shared and Private Spaces

During an initial calibration stage, we estimate each target layer's gradient G_i using training data at the same pretrained model and the same classifier initialization. We then compute

$$\bar{G} = \sum_i p_i G_i, \qquad R_i = G_i - \bar{G}. \tag{3}$$

The mean $\bar{G}$ estimates the gradient of the weighted population objective at the calibration model. The residual R_i measures how client i differs from that objective, with $\sum_i p_i R_i = 0$. We use these

quantities to initialize shared adaptation and select private spaces, respectively. This is an operational allocation of directions; the selected private spaces may overlap directions useful to the population.

Shared initialization. For $\bar{G} = L\Sigma V^\top$, we adopt the disjoint singular-vector blocks of LoRA-GA (Wang et al., 2024a):

$$A_0 = \kappa V_{1:r_g}^\top, \qquad B_0 = \kappa L_{r_g+1:2r_g}. \tag{4}$$

Subscripts select singular-vector columns, and ranks are chosen so that these blocks exist. We retain the scale $\kappa = d_{\text{out}}^{1/4}/\sqrt{\gamma}$, with $\gamma = 16$ in our experiments. This initialization uses different left and right blocks to shape the initial induced update; B_0A_0 is not a paired truncated-SVD reconstruction of $\bar{G}$.

Private space selection. For $R_i = L_i\Sigma_i V_i^\top$, we select paired leading singular directions:

$$B_i^p = \kappa L_{i,1:r_p}, \qquad A_i^p = \kappa V_{i,1:r_p}^\top, \qquad U_{i,0} = 0. \tag{5}$$

The private bases remain frozen throughout training. A full core U_i can combine any selected left direction with any selected right direction; a diagonal core can only rescale matched pairs. This adopts the compact fixed-basis parameterization used by LoRA-XS and LoRA-SB (Bałazy et al., 2024; Ponkshe et al., 2024), with bases selected from client residuals and cores retained locally. Numerically zero residuals produce zero private bases.

We compensate the frozen weight by setting

$$W_f = W_{\text{pre}} - sB_0A_0. \tag{6}$$

Together with $U_{i,0} = 0$, this preserves the pretrained effective weight at initialization in evaluation mode. Calibration and basis selection are performed once; subsequent local updates optimize the task loss within the resulting parameterization.

3.2 Local Adaptation and Federated Aggregation

At round t, the server broadcasts A_{t-1}, B_{t-1}, and the shared classifier. Each participating client restores its private core and fixes the broadcast factors as round anchors. The shared factors start from these anchors, so their increments are initially zero; the private core keeps its previous value. Writing $\delta A_i = A_i - A_{t-1}$ and $\delta B_i = B_i - B_{t-1}$, the local effective weight is

$$\begin{aligned} W_{i,t}^{\text{loc}} = W_f + s\big(&B_{t-1}A_{t-1} + B_{t-1}\delta A_i \\ &+ \delta B_i A_{t-1} + B_i^p U_i A_i^p\big). \end{aligned} \tag{7}$$

The shared increment is the first-order expansion of the factor product around the round anchors. At full-rank anchors, it has a fixed-rank tangent-space interpretation (Vandereycken, 2013; Mishra et al., 2014). Because all clients use the same anchors, averaging factor increments also averages their first-order weight-space effects. The local interaction $\delta B_i \delta A_i$ is omitted.

Clients jointly optimize both shared factor increments, the private core, and the shared classifier using their local losses. Although residual gradients select the private bases, the core is subsequently trained using the actual local loss gradient. It is not updated by explicitly subtracting the population gradient at every step.

For participating clients $\mathcal{S}_t$, let $\widetilde{p}_i = n_i / \sum_{j\in\mathcal{S}_t} n_j$. After local training, the server computes

$$\begin{aligned} \overline{\delta A}_t &= \sum_{i\in\mathcal{S}_t} \widetilde{p}_i \delta A_i, \quad A_t = A_{t-1} + \overline{\delta A}_t, \\ \overline{\delta B}_t &= \sum_{i\in\mathcal{S}_t} \widetilde{p}_i \delta B_i, \quad B_t = B_{t-1} + \overline{\delta B}_t. \end{aligned} \tag{8}$$

Aggregation and evaluation. The classifier is averaged with the same client weights. Private cores remain local, and nonparticipating clients retain their previous cores. Evaluation uses Eq. (2). Expanding the reconstructed shared product gives

$$\begin{aligned} B_tA_t = B_{t-1}A_{t-1} &+ B_{t-1}\overline{\delta A}_t + \overline{\delta B}_t A_{t-1} \\ &+ \overline{\delta B}_t \overline{\delta A}_t. \end{aligned} \tag{9}$$

The middle two terms equal the weighted average of the clients' first-order shared effects. The final term reintroduces a second-order contribution when the server reconstructs the factor product. Thus, linearization applies to the local shared model; server reanchoring is retained in the analysis.

Parameter and communication costs. For each adapted layer, the shared factors contain $r_g(d_{\text{in}} + d_{\text{out}})$ trainable entries. Personalization adds r_p^2 trainable core entries and $r_p(d_{\text{in}} + d_{\text{out}})$ frozen basis entries per client. In each communication round, a participating client receives and uploads the shared factors or their equivalent increments, together with the shared classifier; private bases and cores are not included in these round-wise messages. Initial gradient calibration and any required basis distribution are accounted for separately. These are parameter-entry counts; storage and communication in bytes also depend on numerical precision.

3.3 Theoretical Properties

We now examine the private-space selection criterion: how much calibration residual the chosen space represents, and when this capture implies an attainable improvement in the client objective. We analyze one adapted weight matrix with the remaining model parameters held fixed. Let $F(W) = \sum_i p_i f_i(W)$, where $p_i \geq 0$ and $\sum_i p_i = 1$, and define

$$R_i(W) = \nabla_W f_i(W) - \nabla_W F(W), \qquad R_i^0 = R_i(W_0),$$

where W_0 is the common calibration reference. For an active private branch, write $B_i^p = \kappa Q_{L,i}$ and $A_i^p = \kappa Q_{R,i}^\top$, with $Q_{L,i}^\top Q_{L,i} = Q_{R,i}^\top Q_{R,i} = I_{r_p}$. The unconstrained full core spans

$$\begin{aligned} \mathcal{S}_i &= \left\{ Q_{L,i} C Q_{R,i}^\top : C \in \mathbb{R}^{r_p \times r_p} \right\}, \\ \mathcal{P}_i(X) &= Q_{L,i} Q_{L,i}^\top X Q_{R,i} Q_{R,i}^\top. \end{aligned} \tag{10}$$

Here $\mathcal{P}_i$ is the Frobenius-orthogonal projector onto $\mathcal{S}_i$. The nonzero scale $s\kappa^2$ is absorbed into C. For the zero-residual branch, we instead use $\mathcal{S}_i = \{0\}$ and $\mathcal{P}_i = 0$.

Residual representation. If the bases are exact paired leading singular vectors of R_i^0, then $\mathcal{P}_i(R_i^0)$ is a best approximation of rank at most r_p. By the classical truncated-SVD approximation result (Eckart & Young, 1936),

$$\begin{aligned} \left\| R_i^0 - \mathcal{P}_i(R_i^0) \right\|_F^2 &= \sum_{j > r_p} \sigma_j(R_i^0)^2, \\ \left\| \mathcal{P}_i(R_i^0) \right\|_F^2 &= \sum_{j \leq r_p} \sigma_j(R_i^0)^2. \end{aligned} \tag{11}$$

This is a representation guarantee at calibration; its proof appears in Appendix C.3. Estimated gradients and randomized SVD (Halko et al., 2011) introduce errors relative to this ideal construction; Appendix C.4 bounds their effects.

Conditional value of private adaptation. Consider quadratic client losses in the effective weight coordinates, with vectorized Hessians satisfying $\mu I \preceq H_i \preceq LI$, where $0 < \mu \leq L$. At a common reference $W_\star$ satisfying $\nabla_W F(W_\star) = 0$, set $R_i^\star = R_i(W_\star) = \nabla_W f_i(W_\star)$. Keep the calibration-selected space $\mathcal{S}_i$ fixed and define

$$I_i = f_i(W_\star) - \min_{D \in \mathcal{S}_i} f_i(W_\star + D). \tag{12}$$

Then

$$\frac{\|\mathcal{P}_i(R_i^\star)\|_F^2}{2L} \leq I_i \leq \frac{\|\mathcal{P}_i(R_i^\star)\|_F^2}{2\mu}. \tag{13}$$

Theorem 12 in Appendix N.3 establishes this bound. Thus nonzero projected residual implies a strictly positive attainable improvement under these assumptions. The reference is stationary in effective weight coordinates; stationarity of the shared factors alone does not imply this condition. The result concerns unrestricted optimization within the fixed full-core space, rather than the gain achieved by a finite sequence of local SGD steps.

Connection to initialization. Calibration need not be stationary. For a fixed projector, let $\varepsilon_i^{\text{drift}} = \|R_i(W_\star) - R_i(W_0)\|_F$. Projection contraction and the reverse triangle inequality give

$$I_i \geq \frac{\left[\|\mathcal{P}_i(R_i^0)\|_F - \varepsilon_i^{\text{drift}} \right]_+^2}{2L}, \qquad [x]_+ = \max\{x, 0\}. \tag{14}$$

Table 1: Accuracy (%) on GLUE benchmark protocol (Guo et al., 2025). MNLI-m and MNLI-mm denote the matched and mismatched evaluation sets. Reference rows retain their published means and standard deviations. Avg. is recomputed as the unweighted mean of the five displayed mean scores. Bold marks the highest mean within each adapter family.

	Method	MNLI-m	MNLI-mm	SST-2	QNLI	QQP	Avg.
LoRA	LoRA	$88.71_{\pm 0.09}$	$88.19_{\pm 0.02}$	$95.16_{\pm 0.09}$	$91.16_{\pm 0.72}$	$85.33_{\pm 1.33}$	89.71
	FFA-LoRA	$88.83_{\pm 0.02}$	$88.27_{\pm 0.03}$	$94.95_{\pm 0.04}$	$91.52_{\pm 0.59}$	$86.71_{\pm 0.07}$	90.06
	FedDPA-LoRA	$88.99_{\pm 0.06}$	$88.43_{\pm 0.05}$	$95.50_{\pm 0.06}$	$90.74_{\pm 1.38}$	$85.73_{\pm 1.73}$	89.88
	FedSA-LoRA	$90.18_{\pm 0.02}$	$88.88_{\pm 0.02}$	$96.00_{\pm 0.04}$	$92.13_{\pm 0.24}$	$87.48_{\pm 0.22}$	90.93
	GReP-LoRA (Ours)	$\mathbf{91.54}_{\pm 0.06}$	$\mathbf{91.49}_{\pm 0.02}$	$\mathbf{96.44}_{\pm 0.05}$	$\mathbf{94.16}_{\pm 0.07}$	$\mathbf{89.72}_{\pm 0.09}$	**92.67**
rsLoRA	rsLoRA	$88.91_{\pm 0.15}$	$88.33_{\pm 0.04}$	$95.02_{\pm 0.24}$	$91.21_{\pm 0.39}$	$86.73_{\pm 0.98}$	90.04
	FFA-rsLoRA	$89.21_{\pm 0.11}$	$88.45_{\pm 0.08}$	$95.42_{\pm 0.17}$	$91.42_{\pm 0.44}$	$86.93_{\pm 1.18}$	90.29
	FedDPA-rsLoRA	$89.34_{\pm 0.11}$	$88.53_{\pm 0.05}$	$95.56_{\pm 0.21}$	$90.97_{\pm 0.76}$	$86.81_{\pm 0.53}$	90.24
	FedSA-rsLoRA	$90.35_{\pm 0.11}$	$89.02_{\pm 0.03}$	$95.78_{\pm 0.08}$	$92.03_{\pm 0.22}$	$87.97_{\pm 0.16}$	91.03
	GReP-LoRA (Ours)	$\mathbf{90.50}_{\pm 0.14}$	$\mathbf{89.95}_{\pm 0.04}$	$\mathbf{96.55}_{\pm 0.16}$	$\mathbf{92.22}_{\pm 0.24}$	$\mathbf{89.75}_{\pm 0.51}$	**91.79**
VeRA	VeRA	$85.54_{\pm 0.10}$	$85.09_{\pm 0.07}$	$93.53_{\pm 0.13}$	$91.90_{\pm 0.17}$	$82.07_{\pm 0.35}$	87.63
	FFA-VeRA	$86.63_{\pm 0.13}$	$86.22_{\pm 0.08}$	$93.44_{\pm 0.05}$	$92.05_{\pm 0.23}$	$82.23_{\pm 0.07}$	88.11
	FedDPA-VeRA	$86.74_{\pm 0.11}$	$86.35_{\pm 0.07}$	$93.61_{\pm 0.32}$	$90.73_{\pm 0.54}$	$82.11_{\pm 0.41}$	87.91
	FedSA-VeRA	$87.21_{\pm 0.10}$	$86.52_{\pm 0.04}$	$93.68_{\pm 0.07}$	$\mathbf{92.91}_{\pm 0.09}$	$82.56_{\pm 0.05}$	88.58
	GReP-LoRA (Ours)	$\mathbf{88.23}_{\pm 0.23}$	$\mathbf{88.58}_{\pm 0.83}$	$\mathbf{94.95}_{\pm 0.41}$	$91.84_{\pm 0.28}$	$\mathbf{83.97}_{\pm 0.11}$	**89.51**

Table 2: ViT results on DomainNet and NICO++ with 6 clients and feature heterogeneity.

Method	**Clipart**	**Infograph**	**Painting**	**Quickdraw**	**Real**	**Sketch**	**Average**
Centralized	85.20 ± 0.018	57.15 ± 0.037	81.48 ± 0.014	73.09 ± 0.005	90.90 ± 0.003	78.81 ± 0.036	77.77
FFA-LoRA	81.75 ± 0.038	51.96 ± 0.058	77.51 ± 0.029	61.83 ± 0.095	88.68 ± 0.011	75.20 ± 0.050	72.82
FedIT	84.37 ± 0.069	54.17 ± 0.127	79.67 ± 0.047	69.00 ± 0.085	89.20 ± 0.012	78.08 ± 0.035	75.75
FLoRA	83.70 ± 0.041	53.51 ± 0.075	79.43 ± 0.046	70.09 ± 0.046	89.25 ± 0.011	77.20 ± 0.060	75.53
FlexLoRA	85.15 ± 0.034	53.93 ± 0.132	79.82 ± 0.034	70.01 ± 0.058	89.42 ± 0.010	77.85 ± 0.048	76.02
LoRA-FAIR	86.25 ± 0.032	56.26 ± 0.062	80.09 ± 0.072	71.25 ± 0.039	89.52 ± 0.014	79.06 ± 0.061	77.07
FedSA-LoRA	85.01 ± 0.044	54.20 ± 0.083	**83.02 ± 0.027**	71.05 ± 0.051	91.35 ± 0.031	**81.46 ± 0.020**	77.68
GReP-LoRA (Ours)	**86.46 ± 0.026**	**57.59 ± 0.042**	82.23 ± 0.085	**75.83 ± 0.041**	**91.69 ± 0.039**	80.96 ± 0.027	**79.81**
Method	**Autumn**	**Dim**	**Grass**	**Outdoor**	**Rock**	**Water**	**Average**
Centralized	92.74 ± 0.063	89.63 ± 0.059	93.93 ± 0.024	91.07 ± 0.074	90.96 ± 0.036	90.71 ± 0.054	91.51
FFA-LoRA	91.26 ± 0.019	88.19 ± 0.053	93.29 ± 0.012	89.84 ± 0.024	90.51 ± 0.019	88.60 ± 0.048	90.28
FedIT	91.64 ± 0.024	88.87 ± 0.047	93.09 ± 0.015	90.05 ± 0.028	90.87 ± 0.075	88.96 ± 0.029	90.58
FLoRA	91.48 ± 0.043	89.47 ± 0.063	93.33 ± 0.037	90.38 ± 0.040	90.83 ± 0.041	90.05 ± 0.057	90.93
FlexLoRA	91.26 ± 0.065	88.91 ± 0.042	93.16 ± 0.013	90.41 ± 0.026	90.78 ± 0.029	89.09 ± 0.043	90.60
LoRA-FAIR	92.47 ± 0.032	89.35 ± 0.054	93.73 ± 0.016	90.56 ± 0.025	91.01 ± 0.060	90.34 ± 0.035	91.24
FedSA-LoRA	93.24 ± 0.026	89.99 ± 0.047	94.00 ± 0.019	89.99 ± 0.058	90.47 ± 0.037	90.03 ± 0.012	91.29
GReP-LoRA (Ours)	**94.34 ± 0.025**	**90.75 ± 0.035**	**94.84 ± 0.034**	**91.70 ± 0.016**	**92.41 ± 0.056**	**90.99 ± 0.032**	**92.50**

Corollary 13 in Appendix N.3 proves this connection between initial residual capture and attainable improvement when drift is sufficiently small. Appendix N.5 further accounts for sampling error and residual drift away from calibration. At a nonstationary reference, the corresponding quadratic improvement bound uses the projected actual gradient $\mathcal{P}_i(\nabla_W f_i)$. Appendices N.7 and F analyze actual factor/core updates, server reanchoring, and nonconvex stationarity with explicit stochastic and direction errors.

4 Experiments

We evaluate whether gradient-guided allocation of shared and private adaptation improves federated fine-tuning across vision and language tasks. We then examine whether the private branch specializes to its client and which choices in its construction contribute to performance.

Table 3: ViT results on DomainNet and NICO++ with 30 clients and feature and label heterogeneity.

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
Centralized	85.20 ± 0.018	57.15 ± 0.037	81.48 ± 0.014	73.09 ± 0.005	90.90 ± 0.003	78.81 ± 0.036	77.77
FFA-LoRA	81.75 ± 0.018	51.96 ± 0.022	77.51 ± 0.051	61.83 ± 0.025	88.68 ± 0.106	75.20 ± 0.091	72.82
FedIT	84.08 ± 0.029	52.94 ± 0.024	79.62 ± 0.067	61.03 ± 0.019	88.94 ± 0.074	76.70 ± 0.125	73.89
FLoRA	83.97 ± 0.042	53.57 ± 0.043	80.01 ± 0.063	62.77 ± 0.058	88.95 ± 0.060	76.30 ± 0.082	74.26
FlexLoRA	84.29 ± 0.018	53.60 ± 0.036	79.54 ± 0.084	62.05 ± 0.079	89.23 ± 0.055	76.76 ± 0.085	74.25
LoRA-FAIR	**84.99 ± 0.024**	**55.15 ± 0.058**	**80.51 ± 0.038**	62.77 ± 0.059	89.48 ± 0.026	77.03 ± 0.053	74.99
FedSA-LoRA	83.41 ± 0.063	53.42 ± 0.018	79.37 ± 0.029	73.79 ± 0.037	**90.55 ± 0.043**	**78.56 ± 0.024**	76.52
GReP-LoRA (Ours)	83.84 ± 0.058	53.51 ± 0.045	80.09 ± 0.047	**75.63 ± 0.016**	90.39 ± 0.042	77.74 ± 0.030	**76.84**
Method	Autumn	Dim	Grass	Outdoor	Rock	Water	Average
Centralized	92.74 ± 0.063	89.63 ± 0.059	93.93 ± 0.024	91.07 ± 0.074	90.96 ± 0.036	90.71 ± 0.054	91.51
FFA-LoRA	91.42 ± 0.013	86.99 ± 0.056	92.06 ± 0.045	88.83 ± 0.048	90.10 ± 0.065	87.29 ± 0.035	89.45
FedIT	91.31 ± 0.035	86.91 ± 0.029	92.33 ± 0.008	89.01 ± 0.093	89.97 ± 0.042	87.37 ± 0.068	89.48
FLoRA	91.28 ± 0.105	87.07 ± 0.062	92.27 ± 0.020	89.52 ± 0.064	90.04 ± 0.056	87.43 ± 0.019	89.60
FlexLoRA	91.81 ± 0.031	87.23 ± 0.046	92.45 ± 0.007	89.25 ± 0.021	89.79 ± 0.033	87.37 ± 0.071	89.65
LoRA-FAIR	91.79 ± 0.040	87.59 ± 0.060	92.90 ± 0.016	89.98 ± 0.028	90.39 ± 0.025	87.60 ± 0.018	90.04
FedSA-LoRA	**93.46 ± 0.066**	**89.51 ± 0.016**	93.14 ± 0.043	90.02 ± 0.060	**90.56 ± 0.019**	89.72 ± 0.071	91.07
GReP-LoRA (Ours)	92.47 ± 0.051	89.32 ± 0.038	**93.58 ± 0.025**	**90.74 ± 0.056**	**90.56 ± 0.021**	**90.84 ± 0.057**	**91.27**

4.1 Experimental Setup

Benchmarks and federated settings. For vision, we follow the LoRA-FAIR protocol (Bian et al., 2025) using ViT-B/16 (Dosovitskiy et al., 2021) on the first 100 classes of DomainNet (Peng et al., 2019) and the 60 classes of NICO++ (Zhang et al., 2023). Both datasets contain six domains. The feature-heterogeneous setting assigns one domain to each of six clients. The feature-and-label-heterogeneous setting divides each domain among five clients using a Dirichlet concentration of 0.5, yielding 30 clients. For language, we follow FedSA-LoRA (Guo et al., 2025), evaluating RoBERTa-large (Liu et al., 2019) on MNLI, SST-2, QNLI, and QQP from GLUE (Wang et al., 2019). Data are partitioned among three clients with Dirichlet concentration 0.5; MNLI-m and MNLI-mm denote the two MNLI evaluation splits.

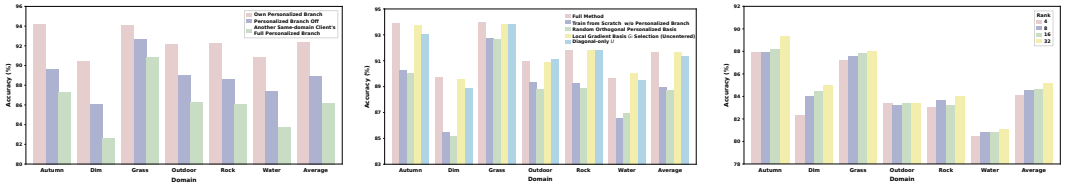
Comparisons and training. Vision baselines include FedIT (Zhang et al., 2024), FFA-LoRA (Sun et al., 2024), FLoRA (Wang et al., 2024b), FlexLoRA (Bai et al., 2024), LoRA-FAIR, and FedSA-LoRA, with centralized fine-tuning as a reference. Language comparisons cover LoRA (Hu et al., 2022), FFA-LoRA, FedDPA-LoRA (Yang et al., 2024), and FedSA-LoRA, together with their rsLoRA (Kalajdzievski, 2023) counterparts. We follow the benchmark training schedules: 50 rounds for DomainNet, 30 for NICO++, and 1,000 for GLUE. The NICO++ ablations are evaluated separately at round 15. We report accuracy in percent and evaluate GReP-LoRA with each client's own private branch. Optimization details and supplementary analyses are provided in Appendix G.

4.2 Main Results

Vision under feature and label heterogeneity. Tables 2 and 3 report the vision results. With six clients, GReP-LoRA obtains reported overall accuracies of 79.81% on DomainNet and 92.50% on NICO++. It improves over LoRA-FAIR in all six domains of both datasets. DomainNet gains include Quickdraw (75.83% versus 71.25%) and Real (91.69% versus 89.52%). On NICO++, it also exceeds FedSA-LoRA in every domain. These results support allocating trainable private capacity within gradient-selected spaces across diverse visual domains.

With 30 clients, the reported overall accuracies are 76.84% and 91.27%, respectively. The gains are less uniform: GReP-LoRA improves over FedSA-LoRA on four DomainNet domains, but is lower on Real and Sketch. On NICO++, it improves on Grass, Outdoor, and Water, ties on Rock, and is lower on Autumn and Dim. Thus, the method remains competitive under combined feature and label heterogeneity without improving every domain. The six- and thirty-client benchmarks also change the label partition and participation regime, so their difference is not an isolated client-count effect.

Language understanding. Table 1 shows that GReP-LoRA improves over FedSA-LoRA on all five reported evaluation columns in the LoRA setting. The gains are 1.36, 2.61, 0.44, 2.03, and 2.24



(a) Private-branch interventions. (b) Component ablations. (c) Joint rank sensitivity.

Figure 3: **Ablation and sensitivity studies on NICO++.** (a) Retaining, disabling, or replacing the private branch. (b) Alternative private bases and core parameterizations. (c) Jointly varying the shared and private ranks, $r_g = r_p \in \{4, 8, 16, 32\}$. Each study uses its own reference configuration.

percentage points on MNLI-m, MNLI-mm, SST-2, QNLI, and QQP, respectively. Improvements also hold across the reported rsLoRA results, including 0.93 points on MNLI-mm and 1.78 points on QQP. This extends the empirical benefit beyond the vision benchmarks and standard LoRA scaling. The VeRA (Kopiczko et al., 2024) results improve on four of the five evaluation columns, while QNLI is lower than FedSA-VeRA; Appendix G.3 provides the task-level comparison.

4.3 Ablations and Further Analysis

All analyses below use NICO++ at communication round 15. We distinguish interventions on a trained model from variants trained with a modified private branch, since these answer different questions. Each study is interpreted against its own reference configuration.

Does the private branch specialize to its client? Figure 3a compares each client's own private branch against disabling it or replacing it with another same-domain client's complete branch. The client's own branch performs best in domain; disabling it lowers accuracy, and replacing it further reduces it. These results suggest client- and trajectory-specific corrections, supporting local retention of the complete branch without isolating the contributions of the core and client-specific bases.

Which private directions are useful? Figure 3b shows that training without a private branch lowers accuracy across all domains. Random orthogonal bases also underperform, suggesting that additional private parameters alone do not explain the gains. Bases from the uncentered local gradient G_i perform similarly to residual-selected bases on average, with domain-dependent differences. These results support gradient-informed private directions, but not a large accuracy advantage from residual centering alone.

How much core flexibility is needed? Restricting U_i to a diagonal matrix retains most of the complete method's performance, with a modest drop in average accuracy. The full core enables interactions between selected left and right directions, whereas a diagonal core only rescales paired directions. Its benefit is smaller than the gap between gradient-selected and random bases. Since the diagonal restriction also reduces active private parameters, this comparison combines expressivity and capacity effects. Appendix G.2 examines joint rank settings $r_g = r_p \in \{4, 8, 16, 32\}$.

5 Conclusion

We revisited how shared and client-specific adaptation should be allocated in federated LoRA. Our diagnostics show that initialization can reverse the apparent asymmetry between the two factors, making factor similarity alone insufficient to determine their roles. Motivated by this observation, we introduced GReP-LoRA, which uses the mean client gradient to initialize shared adaptation and client gradient residuals to select private spaces. The method combines a round-anchored first-order shared model with locally trained cores in frozen private bases. Our analysis connects residual representation to attainable local objective improvement under quadratic losses and bounded curvature. Experiments on DomainNet, NICO++, and GLUE support the effectiveness of this design, while ablations highlight the value of gradient-informed private directions and retaining each client's own private branch. Together, these findings motivate treating sharing and personalization as choices of adaptation spaces grounded in client objectives. Future work includes refreshing private spaces as training progresses and reducing the cost of initial gradient calibration.

AI USE STATEMENT

Generative AI (OpenAI Codex) assisted with literature discovery, interpretation of experiment logs, refinement of the method description and hypotheses, formulation of mathematical claims and proof drafts, numerical checks, plotting code, manuscript writing, and template conversion. Automated numerical checks supplement rather than replace the proofs. The authors are responsible for the mathematical arguments, source attributions, and experimental claims.

REPRODUCIBILITY STATEMENT

The appendix provides full theoretical statements and proofs, experimental protocols, implementation details, and checkpoint-specific diagnostics. The accompanying source includes the archived measurements, plotting scripts, and numerical checks for those diagnostics. Their evaluation rounds and single-seed scope are stated alongside the corresponding results.

REFERENCES

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1a134b50202088aa8c595cc99b310e5a-Abstract-Conference.html.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. LoRA-XS: Low-rank adaptation with extremely small number of parameters. *arXiv preprint arXiv:2405.17604*, 2024. URL https://arxiv.org/abs/2405.17604.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. LoRA-FAIR: Federated LoRA fine-tuning with aggregation and initialization refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3737–3746, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Bian_LoRA-FAIR_Federated_LoRA_Fine-Tuning_with_Aggregation_and_Initialization_Refinement_ICCV_2025_paper.html.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f53a37f820d5be5930415d964f4a0187-Abstract-Conference.html.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2): 217–288, 2011. doi: 10.1137/090771806. URL https://arxiv.org/abs/0909.4061.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on LoRA finetuning dynamics. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d4387c37b3b06e55f86eccdb8cd1f829-Abstract-Conference.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023. URL https://arxiv.org/abs/2312.03732.

Othmar Koch and Christian Lubich. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications*, 29(2):434–454, 2007. doi: 10.1137/050639703. URL https://epubs.siam.org/doi/10.1137/050639703.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1b53ad08de383a049e9668a9d0b6a053-Abstract-Conference.html.

Haoyuan Liang, Xinyu Zhang, Shilei Cao, Guowen Li, and Juepeng Zheng. Tta-feddg: Leveraging test-time adaptation to address federated domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 18658–18666, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db36f4d603cc9e3a2a5e10b93e6428f2-Abstract-Conference.html.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29:591–621, 2014. doi: 10.1007/s00180-013-0464-z. URL https://arxiv.org/abs/1209.0430.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1812.01754.

Kaustubh Ponkshe, Raghav Singhal, Eduard Gorbunov, Alexey Tumanov, Samuel Horvath, and Praneeth Vepakomma. Initialization using update approximation is a silver bullet for extremely efficient low-rank fine-tuning. *arXiv preprint arXiv:2411.19557*, 2024. URL https://arxiv.org/abs/2411.19557.

Raghav Singhal, Kaustubh Ponkshe, Rohit Vartak, Lav R. Varshney, and Praneeth Vepakomma. Fed-SB: A silver bullet for extreme communication efficiency and performance in (private) federated LoRA fine-tuning. *arXiv preprint arXiv:2502.15436*, 2025a. URL https://arxiv.org/abs/2502.15436.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1316–1336, 2025b. doi: 10.18653/v1/2025.acl-long.67. URL https://aclanthology.org/2025.acl-long.67/.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4e243e95c913b367775d71d7182b99d9-Abstract-Conference.html.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. URL https://arxiv.org/abs/1209.3834.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/62c4718cc334f6a0a62fb81c4a2095a1-Abstract-Conference.html.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. FLoRA: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28312c9491d60ed0c77f7fff4ad86dd1-Abstract-Conference.html.

Xinyi Xu, Bingnan Xiao, Shuang Qin, Gang Feng, and Tony Q. S. Quek. Rethinking factor sharing in federated LoRA: A rank-aware adaptive approach. *arXiv preprint arXiv:2608.09742*, 2026. URL https://arxiv.org/abs/2608.09742.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-personalizing adapter for federated foundation models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/45a30141c6719e9cfedfb51f1c665a37-Abstract-Conference.html.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the FederatedGPT: Federated instruction tuning. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6915–6919, 2024. doi: 10.1109/ICASSP48485.2024.10447454. URL https://doi.org/10.1109/ICASSP48485.2024.10447454.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. NICO++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16036–16047, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_NICO_Towards_Better_Benchmarking_for_Domain_Generalization_CVPR_2023_paper.html.

Zihan Zhou, Yang Zhou, Tianshi Che, Zeru Zhang, Jiaxiang Ren, Da Yan, Zhe Jiang, Yelong Shen, Ruoming Jin, and Jianfeng Gao. Noise-free dynamic rank-adaptation via riemannian methods in federated fine-tuning. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300, pp. 478–486, 2026. URL https://proceedings.mlr.press/v300/zhou26a.html.

Zihan Zhu, Zhehang Du, Xuyang Chen, Tim Tsz-Kit Lau, Jiayuan Wu, X. Y. Han, Qi Long, and Weijie Su. Rank-efficient LoRA via joint tangent-space optimization under isotropic curvature. *arXiv preprint arXiv:2609.12123*, 2026. URL https://arxiv.org/abs/2609.12123.

APPENDIX CONTENTS

A Notation and Scope of the Theoretical Claims **15**

B Initialization-Induced Similarity **15**

B.1 A common anchor can dominate cosine similarity 15

B.2 An explicit counterexample to the knowledge interpretation 15

B.3 First-step asymmetry and regularization 15

B.4 Common history and gauge 16

C Mean and Residual Gradient Geometry **16**

C.1 Weighted decomposition 16

C.2 Why the shared initializer uses disjoint index sets 16

C.3 Exact residual representation 17

C.4 Perturbed or approximate spectral bases 17

D Conditional Optimization Advantage **18**

D.1 Nested spaces under an isotropic quadratic 18

D.2 A residual-specific sufficient condition for strict gain 18

D.3 A conditional contraction bound for smooth nonquadratic losses 19

E The Optimizer Is Not the Orthogonal Projector **19**

F Nonconvex Convergence Proof **20**

F.1 Statement and parameter-space assumptions 20

F.2 The weighted gradient 20

F.3 Proof of the direction-error bound in Theorem 6 21

F.4 Bounding local-SGD error for the stated idealization 21

F.5 Server reanchoring and second-order discrepancies 21

F.6 Why the actual training settings need additional assumptions 22

G Experimental Protocols and Additional Analyses **23**

G.1 Benchmark protocols 23

G.2 Joint shared and private rank sensitivity 23

G.3 Adapter-family comparisons 23

G.4 Scope of the component controls 23

H Implementation and Archived Diagnostic Configuration **24**

H.1 Algorithm in executable order 24

H.2 Configuration of the archived vision diagnostics 24

H.3 Parameter and calibration accounting 25

H.4 Forward, dropout and rank qualifications 25

I Extended initialization diagnostics **26**

J Checkpoint-specific accuracy diagnostics **27**

K Why residual gradients, and what does the private core learn? **28**

K.1 Residual gradients differentiate a client from the population objective 28

K.2 When a residual identifies the personalized optimum 28

K.3 The core is a coefficient matrix in a tensor-product basis 29

K.4 Conditional convergence of the private core 30

K.5 Distinguishing stable residual signal from sampling noise 31

L Measured evidence for residual selection and private-core mixing **32**

L.1 Do residual directions predict disjoint training data? 32

L.2 Which components of the learned core affect predictions? 33

L.3 Does a private branch preferentially serve its owner? 34

L.4 Matched training controls at a fixed checkpoint 35

M From factor increments to intrinsic update geometry **36**

M.1 The represented increment and the factorization are different objects 36

M.2 Three orthogonal tangent components 36

M.3 Intrinsic aggregation requires an explicit matrix-space update 37

N Detailed analysis: from similar factors to sharing decisions **38**

N.1 Perfect factor agreement with opposite effective updates 38

N.2 Why mean and residual are the relevant orthogonal split 38

N.3 Conditional objective value of a fixed private space 39

N.4 Calibration away from a common optimum: curvature mismatch 40

N.5 Sampling error and the lifetime of a frozen residual basis 41

N.6 Why a full core can help, and why SVD alone does not prove it 41

N.7 The actual one-step factor/core update 41

N.8 Momentum, local step counts and participation in the convergence bound 43

A Notation and Scope of the Theoretical Claims

All matrix norms are Frobenius norms unless another norm is shown. The inner product is $\langle X, Y\rangle = \operatorname{tr}(X^\top Y)$. Client weights satisfy $p_i > 0$ and $\sum_i p_i = 1$; clients with zero weight can be omitted. We assume the implementation scales $s > 0$ and $\kappa > 0$. Vectorization stacks matrix columns. The product-space inner product is $\langle X, Y\rangle_p = \sum_i p_i \langle X_i, Y_i\rangle$. Layerwise results extend to multiple layers by taking direct sums. The analysis of representation and the isotropic quadratic uses effective weight coordinates, whereas Theorem 6 uses trainable factor, classifier, and private-core coordinates. These two spaces are not interchanged.

Our claims have four different scopes.

1. The initialization, mean-residual, and reanchoring identities are exact algebraic statements, independent of the network's nonconvexity.
2. Exact-SVD residual optimality concerns approximation in a fixed calibration gradient space. Randomized SVD and finite-sample errors can weaken it.
3. The nonconvex convergence theorem bounds average stationarity under explicit smoothness, noise, and direction-error assumptions. It does not prove convergence to a global minimizer of ViT training loss.
4. Objective comparisons use stated quadratic curvature assumptions or a same-start local smoothness bound. They are not unconditional test-accuracy or factor-SGD acceleration theorems.

B Initialization-Induced Similarity

B.1 A common anchor can dominate cosine similarity

For nonzero vectors or vectorized matrices, write $c(x, y) = \langle x, y\rangle / (\|x\| \, \|y\|)$.

Proposition 1 (Similarity near a common anchor)**.** *Let $c \neq 0$ and $0 \leq \epsilon < 1$. If $\|x_i - c\| \leq \epsilon \|c\|$ for each client i, then $x_i \neq 0$ and $c(x_i, x_j) \geq 1 - 2\epsilon^2$ for every pair. This bound concerns complete vectors, irrespective of the directions of their displacements from c.*

Proof. Since $\|x_i - c\| < \|c\|$, the vector x_i is nonzero and $\langle x_i, c\rangle > 0$. Let $\theta_i \in [0, \pi/2)$ be the angle from c to x_i. The distance from c to the line spanned by x_i is $\|c\| \sin\theta_i$. Because x_i is a point on that line,

$$\|c\| \sin\theta_i \leq \|c - x_i\| \leq \epsilon \|c\| .$$

Therefore $\theta_i \leq \arcsin\epsilon$. The angle between x_i and x_j is at most $\theta_i + \theta_j \leq 2\arcsin\epsilon \leq \pi$. Since cosine is decreasing on $[0, \pi]$,

$$c(x_i, x_j) \geq \cos(2\arcsin\epsilon) = 1 - 2\epsilon^2.$$

For $\epsilon = 0$, all vectors equal c and the same statement holds. □

B.2 An explicit counterexample to the knowledge interpretation

Take $c \perp e$ with $\|c\| = 1$ and $\|e\| = \epsilon > 0$. For $x_1 = c + e$ and $x_2 = c - e$,

$$c(x_1, x_2) = \frac{1 - \epsilon^2}{1 + \epsilon^2} \longrightarrow 1 \quad \text{as } \epsilon \longrightarrow 0, \qquad c(e, -e) = -1.$$

Thus arbitrarily high complete-vector similarity is compatible with opposite changes. This establishes non-identifiability from that statistic, without assuming that real client updates must be opposite.

B.3 First-step asymmetry and regularization

For $W = W_f + sBA$, the differential is $dW = s(dB\, A + B\, dA)$, hence

$$df_i = \langle G_i, dW\rangle = \left\langle sG_iA^\top, dB\right\rangle + \left\langle sB^\top G_i, dA\right\rangle .$$

Consequently,

$$\nabla_A f_i = sB^\top G_i, \qquad \nabla_B f_i = sG_i A^\top. \tag{15}$$

With $B_0 = 0$ and a first plain-SGD step of size η, the task-driven updates are $\delta A_i = 0$ and $\delta B_i = -\eta s G_i A_0^\top$. With $A_0 = 0$, the roles reverse. In the reported SGD, coupled weight decay contributes $\lambda_{\mathrm{wd}} A_0$ or $\lambda_{\mathrm{wd}} B_0$. Consequently, the proposition about task gradients does not imply literal freezing under the optimizer. Decay-corrected diagnostics are archived as a supplementary control.

B.4 Common history and gauge

Writing the current local change as $\delta A_{i,t} = A_{i,t} - A_{t-1}$ gives

$$A_{i,t} - A_0 = (A_{t-1} - A_0) + \delta A_{i,t}. \tag{16}$$

The shared vector $A_{t-1} - A_0$ plays the role of c in the same proof. Subtracting initialization therefore removes only one possible anchor. For the gauge example, take nonzero rank-one factors (b, a) and another client's factors $(-b, -a)$. Their products are identical, while both corresponding-factor cosines are -1. This does not imply that such a gauge transformation actually occurred in the archived runs; it shows why factor similarity is not a gauge-invariant semantic quantity.

C Mean and Residual Gradient Geometry

C.1 Weighted decomposition

Expanding $G_i = \bar{G} + R_i$ gives

$$\sum_i p_i \|G_i\|^2 = \left\|\bar{G}\right\|^2 + 2\left\langle \bar{G}, \sum_i p_i R_i \right\rangle + \sum_i p_i \|R_i\|^2 .$$

The middle term vanishes because $\sum_i p_i R_i = \sum_i p_i G_i - \bar{G} = 0$. Hence

$$\sum_i p_i \|G_i\|_F^2 = \left\|\bar{G}\right\|_F^2 + \sum_i p_i \|R_i\|_F^2 . \tag{17}$$

This is orthogonality between the common component and the residual tuple in the *weighted client product space*. It does not assert $\langle \bar{G}, R_i \rangle = 0$ for each client, or orthogonality of their selected singular subspaces.

C.2 Why the shared initializer uses disjoint index sets

Set $I_A = \{1, \ldots, r_g\}$ and $I_B = \{r_g + 1, \ldots, 2r_g\}$. For a plain-SGD step at the calibration point with dropout disabled,

$$\begin{aligned}
\delta A &= -\eta s B_0^\top \bar{G}, \\
\delta B &= -\eta s \bar{G} A_0^\top, \\
s(B_0 \delta A + \delta B A_0) &= -\eta s^2 \kappa^2 \left(L_{I_B} L_{I_B}^\top \bar{G} + \bar{G} V_{I_A} V_{I_A}^\top \right) \\
&= -\eta s^2 \kappa^2 \sum_{j=1}^{2r_g} \sigma_j(\bar{G}) L_j V_j^\top .
\end{aligned}$$

The two sums occupy disjoint singular indices. The linearized update therefore equals a scaled best rank-$2r_g$ approximation to $\bar{G}$. This is the gradient-alignment mechanism of LoRA-GA (Wang et al., 2024a), not a new optimality claim about our shared initializer. For a full product update, $s\delta B \delta A$ adds a second-order term. In our local-linear forward the displayed linearized identity holds for a single gradient step, but server factor reanchoring still introduces its own cross term.

C.3 EXACT RESIDUAL REPRESENTATION

Theorem 2 (Residual representation). *Let $1 \le r_p \le \min(d_{\text{out}}, d_{\text{in}})$, and choose orthonormal $Q_{L,i}, Q_{R,i}$ from paired leading singular vectors of a matrix R_i. The Frobenius-orthogonal projector onto $\mathcal{S}_i = \{Q_{L,i} C Q_{R,i}^\top : C \in \mathbb{R}^{r_p \times r_p}\}$ is $\mathcal{P}_i(X) = Q_{L,i} Q_{L,i}^\top X Q_{R,i} Q_{R,i}^\top$. It satisfies $\mathcal{P}_i(R_i) = (R_i)_{r_p}$ and*

$$\min_{D \in \mathcal{S}_i} \|R_i - D\|_F^2 = \min_{\text{rank}(D) \le r_p} \|R_i - D\|_F^2 = \sum_{j > r_p} \sigma_j(R_i)^2. \tag{18}$$

The captured energy is $\sum_{j \le r_p} \sigma_j(R_i)^2$. Multiplication of the bases by a nonzero constant does not change their full-core span. For an exactly zero residual, the implementation's zero space also attains zero error. This is the classical truncated-SVD approximation property (Eckart & Young, 1936), applied to the private parameterization in Eq. (10)*.*

Proof. Write an exact SVD as $R_i = \sum_j \sigma_j \ell_j v_j^\top$. Let $L_r = [\ell_1, \dots, \ell_{r_p}]$ and $V_r = [v_1, \dots, v_{r_p}]$. Orthogonality of the singular vectors gives

$$L_r L_r^\top R_i V_r V_r^\top = L_r \operatorname{diag}(\sigma_1, \dots, \sigma_{r_p}) V_r^\top = (R_i)_{r_p}.$$

The map $\mathcal{P}_i(X) = L_r L_r^\top X V_r V_r^\top$ is idempotent and self-adjoint. For any X and any $Y = L_r C V_r^\top$, $\langle X - \mathcal{P}_i X, Y \rangle = 0$. Thus it is the orthogonal projection onto the private matrix space, and Pythagoras yields

$$\left\|R_i - L_r C V_r^\top\right\|^2 = \|R_i - \mathcal{P}_i R_i\|^2 + \left\|\mathcal{P}_i R_i - L_r C V_r^\top\right\|^2.$$

The minimum is $\sum_{j > r_p} \sigma_j^2$. Scaling both bases by $\kappa \neq 0$ leaves the space unchanged: choose $C^\star = \kappa^{-2} \operatorname{diag}(\sigma_1, \dots, \sigma_{r_p})$.

For completeness, optimality among *all* rank-r_p approximations follows directly. If $\text{rank}(Y) \le r_p$ and P projects onto its column space, then $\|R_i - Y\|^2 \ge \|(I - P)R_i\|^2 = \|R_i\|^2 - \text{tr}(P R_i R_i^\top)$. In an eigenbasis of $R_i R_i^\top$, the diagonal elements of P lie in $[0, 1]$ and sum to at most r_p; the trace is consequently at most $\sum_{j \le r_p} \sigma_j^2$. The truncated SVD achieves the bound. Multiplying the per-client result by p_i and summing proves the energy statement. If $R_i = 0$, the minimum is zero with the specified zero bases. □

Recentering after projection. The truncated components need not have zero weighted mean. Define $H_p = \sum_i p_i \|R_i - \mathcal{P}_i R_i\|_F^2$. If $E_i = R_i - \mathcal{P}_i R_i$ and $\bar{E} = \sum_i p_i E_i$, then

$$\sum_i p_i \left\|E_i - \bar{E}\right\|^2 = \sum_i p_i \|E_i\|^2 - \left\|\bar{E}\right\|^2 \le H_p.$$

This is another approximation identity, not an assertion that the implemented optimizer explicitly recenters private updates.

C.4 PERTURBED OR APPROXIMATE SPECTRAL BASES

Suppose the estimated residual is $\widehat{R}_i$ with $\left\|\widehat{R}_i - R_i\right\| \le \varepsilon_i$. Suppose a matrix $\widetilde{R}_i$ in the chosen private space satisfies $\left\|\widehat{R}_i - \widetilde{R}_i\right\| \le \min_{\text{rank}(Y) \le r_p} \left\|\widehat{R}_i - Y\right\| + \delta_i$. Let $\tau_i = (\sum_{j > r_p} \sigma_j(R_i)^2)^{1/2}$. The triangle inequality, using $(R_i)_{r_p}$ as a candidate approximation to $\widehat{R}_i$, gives

$$\|R_i - \mathcal{P}_i R_i\| \le \left\|R_i - \widetilde{R}_i\right\| \le \tau_i + 2\varepsilon_i + \delta_i.$$

Here δ_i records computational approximation and ε_i can represent calibration sampling error. If Frobenius-orthonormal bases are only approximate numerically, orthonormalize them before using the projector in this formula. Spectral gaps affect basis stability; no gap assumption is needed for the exact best approximation error itself.

D Conditional Optimization Advantage

D.1 Nested spaces under an isotropic quadratic

Let $\mathcal{S}$ be a fixed shared matrix space, such as the tangent space at full-rank anchors, and define $\mathcal{V}_0 = \{(X)_i : X \in \mathcal{S}\}$ and $\mathcal{V}_+ = \mathcal{V}_0 + \prod_i \mathcal{S}_i$. Let Π_0, Π_+ be their orthogonal projectors in the weighted client inner product. Shared and private spaces need not be orthogonal.

Theorem 3 (Conditional finite-step advantage)**.** *For $Q(D) = Q(0) + \langle G, D\rangle_p + \lambda \|D\|_p^2/2$, with $\lambda > 0$, run projected gradient descent from zero in either space with the same $0 < \eta < 2/\lambda$. At iteration $k \geq 0$,*

$$Q(D_0^k) - Q(D_+^k) = \frac{1-(1-\eta\lambda)^{2k}}{2\lambda} \underbrace{\|(\Pi_+ - \Pi_0)G\|_p^2}_{\Gamma} \geq 0. \tag{19}$$

Both excess losses contract by $(1-\eta\lambda)^{2k}$ relative to their respective subspace optima. This is a matrix-space projected-GD result; it does not identify ordinary factor SGD with orthogonal projection.

Proof. For either fixed linear subspace $\mathcal{V}$, let Π be its orthogonal projector. Starting in $\mathcal{V}$, projected gradient descent remains there and obeys

$$D^{k+1} = D^k - \eta\Pi(G + \lambda D^k) = (1-\eta\lambda)D^k - \eta\Pi G.$$

Let $q = 1-\eta\lambda$, so $|q| < 1$. Solving this recurrence gives $D^k = -(1-q^k)\Pi G/\lambda$ and $D^\star = -\Pi G/\lambda$. Completing the square within $\mathcal{V}$ yields

$$Q(D^k) = Q(0) - \frac{1-q^{2k}}{2\lambda}\|\Pi G\|_p^2, \qquad Q(D^k) - Q(D^\star) = \frac{q^{2k}}{2\lambda}\|\Pi G\|_p^2.$$

Because $\mathcal{V}_0 \subseteq \mathcal{V}_+$, $\Pi_+\Pi_0 = \Pi_0\Pi_+ = \Pi_0$. Therefore $\Pi_+ - \Pi_0$ is the orthogonal projector onto $\mathcal{V}_+ \cap \mathcal{V}_0^\perp$ and $\|\Pi_+ G\|_p^2 - \|\Pi_0 G\|_p^2 = \|(\Pi_+ - \Pi_0)G\|_p^2$. Subtracting the two objective formulas proves Eq. (19). The scalar prefactor is positive for $k \geq 1$, so strictness is equivalent to $\Gamma > 0$. □

D.2 A residual-specific sufficient condition for strict gain

The abstract coverage quantity Γ can be connected to the selected residual directions. Let $\mathcal{S}$ be the shared matrix tangent space, so $\mathcal{V}_0 = \{(X, \ldots, X) : X \in \mathcal{S}\}$. Let $P_{\mathcal{S}}$ be its Frobenius projector and define

$$h = \sum_i p_i \|\mathcal{P}_i R_i\|^2, \qquad \epsilon_g = \left\|\bar{G} - P_{\mathcal{S}}\bar{G}\right\|.$$

Corollary 4 (Captured residual energy supplies additional coverage)**.** *For the spaces above,*

$$\Gamma \;\geq\; \left(\max\{0, \sqrt{h} - \epsilon_g\}\right)^2.$$

In particular, $h > \epsilon_g^2$ guarantees the strict advantage in Theorem 3. If $\bar{G} \in \mathcal{S}$, then $\Gamma \geq h$.

Proof. The shared projection of G is the repeated matrix $P_{\mathcal{S}}\bar{G}$. Put $T = (\mathcal{P}_i R_i)_i$. Then $T \in \mathcal{V}_+$, $\|T\|_p^2 = h$, and $\langle R, T\rangle_p = h$. Also $R \perp \mathcal{V}_0$ since $\sum_i p_i R_i = 0$. If $h = 0$, the claimed lower bound is zero and immediate. Otherwise put $v = (I - \Pi_0)T$, which belongs to $\mathcal{V}_+ \cap \mathcal{V}_0^\perp$. We have $\langle R, v\rangle_p = h$, so $v \neq 0$, and $\|v\|_p \leq \sqrt{h}$. Writing E for the repeated matrix $\bar{G} - P_{\mathcal{S}}\bar{G}$,

$$\frac{\langle G, v\rangle_p}{\|v\|_p} = \frac{h + \langle E, v\rangle_p}{\|v\|_p} \geq \frac{h}{\|v\|_p} - \|E\|_p \geq \sqrt{h} - \epsilon_g.$$

The norm of the projection of G onto the extra space is the supremum of its inner product over unit vectors in that space. It is nonnegative and at least the displayed lower bound. Squaring proves the result. □

For exact Ar/B2r bases at calibration, $\epsilon_g^2 = \sum_{j>2r_g} \sigma_j(\bar{G})^2$. For exact residual bases, $h = \sum_i p_i \sum_{j\leq r_p} \sigma_j(R_i)^2$. This gives a measurable, sufficient spectral condition rather than a universal claim. It compares captured residual energy with uncaptured mean-gradient energy; it is a theoretical sufficient condition, not an empirical explanation of all later training rounds.

D.3 A CONDITIONAL CONTRACTION BOUND FOR SMOOTH NONQUADRATIC LOSSES

Corollary 5 (PL condition and gradient coverage)**.** *Let a differentiable loss f be L-smooth and satisfy $\|\nabla f(z)\|^2 \geq 2\mu(f(z) - f_\star)$ along a projected-GD trajectory. Assume its projector obeys $\|\Pi\nabla f(z)\|^2 \geq c\,\|\nabla f(z)\|^2$ for some $c > 0$ on the same trajectory. For $0 < \eta \leq 1/L$,*

$$f(z_k) - f_\star \leq (1 - \eta\mu c)^k (f(z_0) - f_\star).$$

Proof. Smoothness and orthogonality give

$$f(z - \eta\Pi\nabla f(z)) \leq f(z) - \eta(1 - L\eta/2)\,\|\Pi\nabla f(z)\|^2 \leq f(z) - \eta\mu c(f(z) - f_\star).$$

Iteration proves the claim (with the usual compatible constants so the contraction factor lies in $[0, 1)$). □

For nested subspaces, projected gradient energy cannot decrease at a *fixed point*. If uniform constants $c_+ > c_0$ hold over the relevant trajectories with the same L, μ, the certified contraction bound improves. Nestedness alone does not prove these uniform constants or compare losses along two different nonlinear trajectories. The PL property and improved coverage are not established for the experimental ViT.

E THE OPTIMIZER IS NOT THE ORTHOGONAL PROJECTOR

For fixed shared anchors, the linear map from factor increments to a weight perturbation is $\mathcal{J}(\delta A, \delta B) = s(B\delta A + \delta B A)$. Its adjoint is $\mathcal{J}^* G = (sB^\top G, sGA^\top)$. A factor-SGD step therefore induces

$$\delta W = -\eta\mathcal{J}\mathcal{J}^* G = -\eta s^2(BB^\top G + GA^\top A).$$

For orthonormal, equally scaled bases, the tangent-space projector is

$$P_{\mathcal{S}}(G) = P_B G + G P_A - P_B G P_A,$$

where P_B and P_A project onto the column space of B and row space of A. The factor-SGD operator lacks the overlap subtraction and includes scale factors. Away from the initial singular geometry, it cannot be replaced by $P_{\mathcal{S}}$ in a proof.

For the private map $\mathcal{J}_i^p(C) = sB_i^p C A_i^p$,

$$\nabla_{U_i} f_i = s(B_i^p)^\top G_i (A_i^p)^\top, \quad \delta W_i^p = -\eta s^2 B_i^p (B_i^p)^\top G_i (A_i^p)^\top A_i^p.$$

With bases $\kappa L_r, \kappa V_r^\top$, this equals $-\eta s^2 \kappa^4 \mathcal{P}_i(G_i)$. It projects the *current full local gradient*, which can include components of $\bar{G}$, not merely the calibration residual.

With normalized private coordinates and an L_W-smooth weight-space loss, a private-only step $W^+ = W - \eta\mathcal{P}_i(G_i)$ satisfies

$$f_i(W^+) \leq f_i(W) - \eta(1 - L_W\eta/2)\,\|\mathcal{P}_i(G_i)\|^2.$$

This follows from the same smoothness inequality and is a valid local descent claim. Selecting bases to maximize capture of R_i does not necessarily maximize capture of G_i.

Why added directions need not improve the actual rate. In a quadratic with identity Hessian, suppose the original parameter map induces $K_0 = \mathrm{diag}(1, 0)$ and an added weak direction gives $K_+ = \mathrm{diag}(1, \varepsilon)$, $0 < \varepsilon \ll 1$. The attainable optimum can improve, yet the newly reachable error component contracts only by $1 - \eta\varepsilon$ per step. A badly scaled added direction can also increase the largest eigenvalue and require a smaller stable learning rate. This counterexample excludes an unconditional faster-rate claim for the implemented factor/core optimizer.

F NONCONVEX CONVERGENCE PROOF

F.1 STATEMENT AND PARAMETER-SPACE ASSUMPTIONS

Let $x = (\theta, u_1, \dots, u_N)$ contain the shared factors and classifier and all private cores. Define the evaluated endpoint objective $F(x) = \sum_i p_i f_i(\theta, u_i)$ and the weighted parameter norm $\|x\|_P^2 = \|\theta\|^2 + \sum_i p_i \|u_i\|^2$. The gradient $\nabla_P F$ is taken in this inner product. Assume $F \geq F_{\inf}$ and that its gradient is L_F-Lipschitz in this metric on a region containing every analyzed update segment. For the round-start filtration $\mathcal{F}_t$, write the actual parameter update as

$$x_{t+1} = x_t - \alpha(\nabla_P F(x_t) + b_t + \xi_t), \qquad \mathbb{E}_t \xi_t = 0, \qquad \mathbb{E}_t \|\xi_t\|_P^2 \leq v^2. \tag{20}$$

Here α is a round optimization step size, distinct from the LoRA scaling hyperparameter used in s in the method section. Here $\mathbb{E}_t$ is conditional expectation given $\mathcal{F}_t$, b_t is the measurable conditional mean direction error, and v^2 is a uniform conditional second-moment bound. These are assumptions on the round update, not automatic properties of arbitrary client sampling or adaptive optimizers.

Theorem 6 (Stationarity with direction error)**.** *Under these assumptions and $0 < \alpha \leq 1/(4L_F)$, for deterministic x_0,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla_P F(x_t)\|_P^2 \leq \frac{2(F(x_0) - F_{\inf})}{\alpha T} + L_F \alpha v^2 + \frac{4}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|b_t\|_P^2 . \tag{21}$$

The bound explicitly retains persistent direction bias. It does not assert convergence to a global optimum or improved test accuracy.

Corollary 7 (Equal-step, full-participation SGD)**.** *Suppose each client takes $K \geq 1$ plain-SGD steps of size η on its round-anchored surrogate, with full participation. Assume unbiased local gradient estimates bounded by M, surrogate gradients that are $\widetilde{L}$-Lipschitz on the local segments, and agreement between surrogate and evaluated-objective gradients at the round anchor. For $\alpha = K\eta \leq 1/(4L_F)$, $\|b_t\|_P \leq \widetilde{L}M\eta(K-1)/2 \leq \widetilde{L}M\alpha/2$, and consequently*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla_P F(x_t)\|_P^2 \leq \frac{2(F(x_0) - F_{\inf})}{\alpha T} + L_F \alpha v^2 + \widetilde{L}^2 M^2 \alpha^2. \tag{22}$$

With uniform constants and $\alpha = c/\sqrt{T}$, where $0 < c \leq 1/(4L_F)$, this is $O(T^{-1/2})$ average stationarity.

The actual local-linear surrogate and the bilinear endpoint model have equal values and first derivatives at zero shared increments. This is the bridge used here; the endpoint objective retains server reanchoring. Appendix N.8 treats reset momentum, clipping, unequal local step counts, and participation as explicit direction errors. For compact notation in the following proof, write $L = L_F$.

F.2 THE WEIGHTED GRADIENT

Write $g_{\theta,i} = \nabla_\theta f_i$ and $g_{u,i} = \nabla_{u_i} f_i$. By the definition of the weighted inner product,

$$\nabla_P F(x) = \left(\sum_i p_i g_{\theta,i}, g_{u,1}, \dots, g_{u,m}\right).$$

The private components contain no extra factor p_i because that weight is already in the parameter-space metric. Full-participation private SGD using the unweighted local loss is therefore consistent with this gradient convention. The squared stationarity measure is $\|\sum_i p_i g_{\theta,i}\|^2 + \sum_i p_i \|g_{u,i}\|^2$.

F.3 Proof of the direction-error bound in Theorem 6

Let $g_t = \nabla_P F(x_t)$ and let $\mathcal{F}_t$ contain the round-start state and all preceding randomness. Both g_t and b_t are $\mathcal{F}_t$-measurable. Conditional smoothness and $\mathbb{E}_t \xi_t = 0$ give

$$\begin{aligned}\mathbb{E}_t F(x_{t+1}) &\le F(x_t) - \alpha \langle g_t, g_t + b_t \rangle_P + \frac{L\alpha^2}{2}\left(\|g_t + b_t\|_P^2 + \mathbb{E}_t \|\xi_t\|_P^2\right) \\ &\le F(x_t) - \alpha \|g_t\|_P^2 + \alpha\left(\frac{1}{4}\|g_t\|_P^2 + \|b_t\|_P^2\right) + L\alpha^2(\|g_t\|_P^2 + \|b_t\|_P^2) + \frac{L\alpha^2 v^2}{2} \\ &\le F(x_t) - \frac{\alpha}{2}\|g_t\|_P^2 + 2\alpha \|b_t\|_P^2 + \frac{L\alpha^2 v^2}{2}.\end{aligned}$$

The second line uses Young's inequality $|\langle g, b\rangle| \le \|g\|^2/4 + \|b\|^2$ and $\|g+b\|^2 \le 2\|g\|^2 + 2\|b\|^2$. The last line uses $L\alpha \le 1/4$. Take full expectation, sum over t, and use $\mathbb{E}F(x_T) \ge F_{\inf}$. Dividing by $\alpha T/2$ proves Eq. (21). □

F.4 Bounding local-SGD error for the stated idealization

Fix a round t. Write $z_i = (\theta, u_i)$, $z_{i,0} = (\theta_t, u_{i,t})$, and let $\widetilde{f}_{i,t}$ be the round-anchored local-linear surrogate. Use equal local step counts K, full participation, and updates $z_{i,k+1} = z_{i,k} - \eta h_{i,k}$. Assume

$$\mathbb{E}[h_{i,k} \mid z_{i,k}, \mathcal{F}_t, \text{previous local randomness}] = \nabla \widetilde{f}_{i,t}(z_{i,k}), \qquad \|h_{i,k}\| \le M.$$

Assume $\nabla \widetilde{f}_{i,t}$ is $\widetilde{L}$-Lipschitz on these local segments and $\nabla \widetilde{f}_{i,t}(z_{i,0}) = \nabla f_i(z_{i,0})$. The last equality holds for the first-order factor expansion because the removed product and its first derivatives vanish at zero increments. It also holds with the same coupled regularizer and the same stochastic dropout objective; it need not hold if the two objectives are changed inconsistently.

Pathwise, $\|z_{i,k} - z_{i,0}\| \le \eta k M$. The expected mean local direction differs from the true initial gradient by

$$\begin{aligned} e_i &= \mathbb{E}_t\left[\frac{1}{K}\sum_{k=0}^{K-1} h_{i,k}\right] - \nabla f_i(z_{i,0}), \\ \|e_i\| &\le \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_t \left\|\nabla \widetilde{f}_{i,t}(z_{i,k}) - \nabla \widetilde{f}_{i,t}(z_{i,0})\right\| \\ &\le \widetilde{L}\eta M(K-1)/2.\end{aligned}$$

Write $e_i = (e_{\theta,i}, e_{u,i})$. Averaging the shared part and keeping private endpoints produces $b_t = (\sum_i p_i e_{\theta,i}, e_{u,1}, \ldots, e_{u,m})$. Jensen's inequality gives

$$\|b_t\|_P^2 \le \sum_i p_i(\|e_{\theta,i}\|^2 + \|e_{u,i}\|^2) \le \widetilde{L}^2\eta^2 M^2 (K-1)^2/4 \le \widetilde{L}^2 M^2 \alpha^2/4.$$

The concatenated round-noise variance is assumed bounded by v^2; the bounded-gradient assumption also ensures a finite bound. Substitution into Eq. (21) proves Corollary 7. For $\alpha = c/\sqrt{T}$ with $c \le 1/(4L)$, the first and noise terms are $O(T^{-1/2})$, and the drift term is $O(T^{-1})$. A uniformly random round has the same expected squared gradient bound as the average. This does not imply that every individual iterate converges or that a global optimum is reached. □

F.5 Server reanchoring and second-order discrepancies

For common anchors (A, B), direct expansion gives

$$(B + \overline{\delta B})(A + \overline{\delta A}) = BA + B\overline{\delta A} + \overline{\delta B}A + \overline{\delta B}\,\overline{\delta A}.$$

The difference between the last expression and the averaged shared local-linear matrix is exactly the last product. Its norm satisfies

$$\left\|s\overline{\delta B}\,\overline{\delta A}\right\|_F \le s\left\|\overline{\delta B}\right\|_F \left\|\overline{\delta A}\right\|_F \le s D_B D_A$$

if the mean factor increments are bounded by D_B, D_A. It is $O(\eta^2 K^2)$ when both increment norms are $O(\eta K)$. For full local BA, the discrepancy relative to averaging local products is

$$\overline{\delta B}\,\overline{\delta A} - \sum_i \tilde{p}_i \delta B_i \delta A_i = -\sum_i \tilde{p}_i (\delta B_i - \overline{\delta B})(\delta A_i - \overline{\delta A}).$$

By contrast, our local-linear procedure omits the second term on the left locally and retains the first after factor reanchoring. The convergence proof evaluates the smooth *bilinear* objective at round endpoints, so this effect is included through surrogate drift; it is not silently declared zero.

F.6 Why the actual training settings need additional assumptions

Partial participation. Even if clients are sampled uniformly without replacement, normalized sample-size weights need not be unbiased for full-population weights. With two clients and one selected uniformly, $\mathbb{E}[\sum_{i\in\mathcal{S}} \tilde{p}_i g_i] = (g_1 + g_2)/2$, which differs from $p_1 g_1 + p_2 g_2$ when $p_1 \neq p_2$. Private cores are also updated only when selected, yielding a participation-dependent preconditioner. These differences enter b_t under Eq. (20); the vanishing-bias claim does not follow without further weighting or assumptions.

Unequal local steps and optimizer state. Equal local epochs do not imply equal K when client dataset sizes differ. The resulting direction can weight gradients by both sample count and number of steps. Momentum, reset optimizer state, and clipping change the round direction further. A convergence theorem for these exact mechanisms requires controlling their conditional error or analyzing the enlarged momentum state. We do not replace this missing argument with an assertion that the SGD corollary applies unchanged.

Fixed learning rate. The runs use a fixed local learning rate. Even with all the idealized assumptions, the noise and local-drift terms leave a nonzero upper-bound floor at fixed α. The $O(T^{-1/2})$ result describes a specified horizon-dependent diminishing step choice, not the recorded schedule.

From optimization to accuracy. All theorems concern training objectives or gradient norms. No theorem here orders test classification accuracy. A lower approximation floor can motivate the method while overfitting, anisotropic curvature, stale bases, and noise can still reverse its empirical advantage.

G Experimental Protocols and Additional Analyses

G.1 Benchmark protocols

Vision. We follow the LoRA-FAIR benchmark protocol (Bian et al., 2025). The six-client setting uses one domain per client, full participation, and two local epochs per round. The thirty-client setting uses five clients per domain, a within-domain Dirichlet label partition with concentration 0.5, 18 participating clients per round, and five local epochs. Batch size is 128 and the SGD learning rate is 0.01. The benchmark training budgets are 50 communication rounds for DomainNet and 30 for NICO++; the NICO++ ablations in Figure 3 are evaluated at round 15. Supplementary DomainNet diagnostics below retain their separately stated checkpoint rounds, seed, and evaluation subsets.

Language. The GLUE protocol follows FedSA-LoRA (Guo et al., 2025): RoBERTa-large, three clients partitioned with Dirichlet concentration 0.5, batch size 128, ten local minibatch updates per round, and 1,000 communication rounds. The reference LoRA configuration adapts the query and value projections with rank eight and scaling parameter 16. The SGD learning-rate search for LoRA is $\{0.005, 0.01, 0.02, 0.05, 0.1\}$, and for rsLoRA it is $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05\}$. The original VeRA (Kopiczko et al., 2024) baseline configuration uses rank 256 and AdamW with separate learning rates for its adaptation parameters and classifier. These describe the reference benchmark configurations; the rank sweeps and the archived vision configuration are identified separately.

Reported metrics and evaluation. Table 1 contains four GLUE tasks with five evaluation columns because MNLI has matched and mismatched splits. Its Avg. column is the arithmetic mean of those five displayed mean scores; MNLI therefore contributes two evaluation scores to that summary. All task-level means and their accompanying deviations are reported separately. Personalized evaluation uses the evaluated client's own private branch. Removing that branch at a fixed checkpoint is an evaluation intervention, whereas omitting it throughout training produces a different optimization trajectory.

G.2 Joint shared and private rank sensitivity

Figure 3c jointly varies both ranks, $r_g = r_p = r$ for $r \in \{4, 8, 16, 32\}$, on NICO++ at round 15. The reported average accuracy increases over this range, with rank 32 giving the highest average. The response varies across domains: Autumn and Dim improve more visibly, whereas Outdoor changes little and is not monotonic. This observation concerns the tested range and does not establish an optimal rank beyond it.

Increasing both ranks changes both adaptation capacity and resource use. Per adapted layer, the shared factors contain $r(d_{\text{in}} + d_{\text{out}})$ trainable entries, which also determine the size of the shared-factor messages. The private branch adds r^2 trainable core entries and $r(d_{\text{in}} + d_{\text{out}})$ frozen basis entries. The observed trend therefore cannot be attributed solely to private rank, nor obtained at a constant communication budget.

G.3 Adapter-family comparisons

In the rsLoRA setting, GReP-LoRA improves over FedSA-rsLoRA by 0.15, 0.93, 0.77, 0.19, and 1.78 percentage points on MNLI-m, MNLI-mm, SST-2, QNLI, and QQP, respectively. The VeRA-setting results are more mixed. Relative to FedSA-VeRA, GReP-LoRA improves MNLI-m by 1.02 points, MNLI-mm by 2.06, SST-2 by 1.27, and QQP by 1.41, while QNLI is lower by 1.07. The mean of the five displayed VeRA scores is 89.51%, compared with 88.58% for FedSA-VeRA. Thus, the benefit is not uniform across every task and adapter family.

G.4 Scope of the component controls

The NICO++ intervention replaces the complete private branch, including its left basis, core, and right basis. This preserves the donor's coordinate system and tests the utility of its learned correction for the recipient. It does not isolate the core from the bases. The training-time random-basis and uncentered-gradient controls test direction selection. Restricting the core to its diagonal changes both allowed interactions and the number of trainable coefficients, from r_p^2 to r_p. These comparisons do

not independently isolate the contribution of the first-order shared forward. The archived DomainNet controls in Table 6 provide complementary evidence at their stated round-6 checkpoint.

H Implementation and Archived Diagnostic Configuration

H.1 Algorithm in executable order

1. Fix the pretrained backbone, fresh common classifier, random seed, client training splits, and calibration data order. Accumulate each target weight's sample-mean dense gradient for every client without any optimizer step. Compute $\bar{G}$ and R_i.
2. Construct the shared Ar/B2r factors, compensate W_f, and construct paired residual private bases. Set all $U_i = 0$. Save calibration and initialization audits.
3. At each communication round, select the saved client set. Restore the global A, B, and head, plus the selected client's own persistent U_i. Set round anchors to the restored global factors.
4. Optimize Eq. (7) with the configured local epochs. Save factor endpoints, core state, sample count, loss, and optimizer-step count. Do not update private bases.
5. Average the shared factor and head increments by selected-client sample count. Retain each updated core separately; leave absent clients' cores unchanged. Refresh global anchors.
6. At scheduled evaluation rounds, evaluate each assigned test split with its own core. Re-evaluate the same shared checkpoint with all private branches disabled. Save these under separate fields.

H.2 Configuration of the archived vision diagnostics

The configuration below describes the archived seed-51 vision runs used for the supplementary diagnostics. It does not assign a single seed or an identical rank to every experiment in the main result tables.

Item	Configuration
Backbone	ViT-B/16, 12 blocks, width 768, MLP width 3072
Pretraining identifier	`vit_base_patch16_224.augreg2_in21k_ft_in1k`
Adapter targets	Patch embedding projection; each block's attention projection and MLP `fc2`
Image size	224×224
Data transform	ToTensor; bicubic resize, antialias=False; channel normalization
Normalization mean	$(0.48145466, 0.45782750, 0.40821073)$
Normalization standard deviation	$(0.26862954, 0.26130258, 0.27577711)$
Global/private rank	$16/16$ for the archived reference configuration
Adapter scale	$s = 8/16 = 0.5$
Private initialization	Zero 16×16 core, frozen paired residual bases
Optimizer	SGD, learning rate 0.01, momentum 0.9, coupled weight decay 10^{-5}
Gradient clipping	Global trainable-parameter norm threshold 10
Precision	FP32 network; FP64 dense-gradient accumulation and spectral decomposition
Server aggregation	Selected-client sample-size weighted shared factor and head increments
Classifier	One shared trainable classifier; 100 DomainNet or 60 NICO++ outputs
Randomness	Seed 51; saved data splits, client selection, and local data order
Model selection	Fixed final communication round; interim checkpoints labeled separately

The pretrained checkpoint content identifier is

`32aa17d6e17b43500f531d5f6dc9bc93e56ed8841b8a75682e1bb295d722405b.`

The source manifest and run configurations in the accompanying data directory preserve full hashes and method-specific settings. The round-0 effective-weight and evaluation-logit checks concern evaluation mode, not the stochastic dropout output during training.

H.3 Parameter and calibration accounting

The shared adapter factors contain

$$16(768 + 768) + 12 \cdot 16(768 + 768) + 12 \cdot 16(768 + 3072) = 1{,}056{,}768$$

entries. A DomainNet classifier has $768 \cdot 100 + 100 = 76{,}900$ parameters; the shared trainable total is 1,133,668. The proposed client adds 6,400 trainable core entries and 1,056,768 frozen basis entries. These frozen bases are client-dependent, unlike a single reusable global basis.

The target dense weight matrices contain $768 \cdot 768 + 12(768 \cdot 768 + 768 \cdot 3072) = 35{,}979{,}264$ entries. One FP64 gradient bank for these targets alone requires about 274.5 MiB, before temporary buffers and other model tensors. This is a one-time calibration storage/communication consideration, not a per-round private-core upload. DomainNet residual reconstruction reused the archived global gradient and required a further data pass to recover individual client gradients. Total wall-clock cost must include that pass when reporting the actual experimental execution history. A fresh joint calibration can compute both the mean and residual ingredients from the same collection of client gradients.

H.4 Forward, dropout and rank qualifications

Dropout and the initial model. All local adapter terms use the same input-dropout realization. At evaluation dropout is disabled and Eq. (6) is exact up to rounding. With training dropout $\mathcal{D}(x)$, however, the initial layer output is $W_{\mathrm{pre}}x + sB_0A_0(\mathcal{D}(x) - x)$, including a zero-mean perturbation for inverted dropout. This does not imply equality of expected nonlinear network losses. Our idealized weight-space analysis either disables dropout or places its stochastic gradients and calibration mismatch in the stated error terms. The archived vision configuration uses dropout 0.1.

Cost and rank. With 25 modules and $r_p = 16$, the added trainable private budget is $25r_p^2 = 6{,}400$ parameters per client. This excludes 1,056,768 frozen private-basis entries per client, extra forward computation, and dense gradient calibration. Relative to W_f, the evaluated adapter has rank at most $r_g + r_p = 32$; relative to W_{pre} the subtraction in Eq. (6) gives a bound of $2r_g + r_p = 48$. Equal nominal shared rank is not an equal-capacity comparison.

I EXTENDED INITIALIZATION DIAGNOSTICS

The following measurements retain the same vectorwise, within-layer, cross-client definition used in the main text. They distinguish complete factors, cumulative changes, and current-round changes. Missing endpoints are not interpolated into additional observations.

Table 4: Cross-client signed vector cosine. Δ is cumulative from initialization; δ is the current-round change. All values are measured.

Initialization	Round	A	B	ΔA	ΔB	δA	δB
FedIT	1	0.9959	0.2859	0.5229	0.2859	0.5229	0.2859
Reversed FedIT	1	0.1649	0.9998	0.1649	0.2185	0.1649	0.2185
Gradient + BA	1	0.9954	0.9871	0.2230	0.2881	0.2230	0.2881
FedIT	50	0.9959	0.9899	0.9892	0.9899	-0.0480	-0.0694
Reversed FedIT	50	0.9745	0.9999	0.9745	0.9959	-0.0274	-0.0130
Gradient + BA	50	0.9948	0.9960	0.9853	0.9843	-0.0239	-0.0365

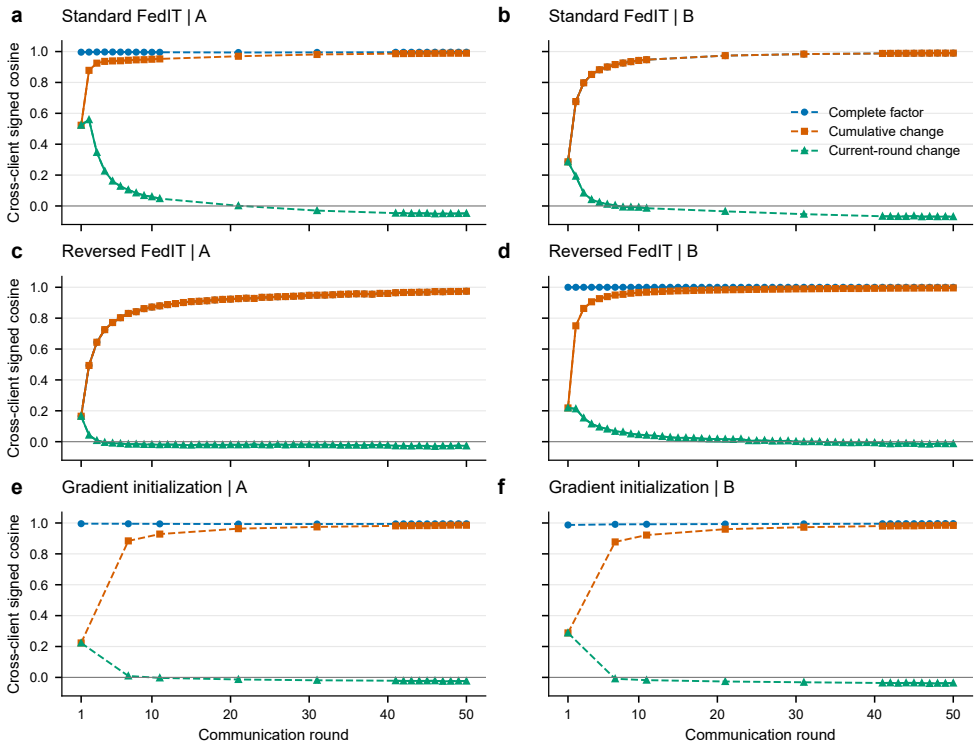


Figure 4: Complete factors, cumulative changes, and current-round changes tell different stories. Markers are archived or deterministically recovered client endpoints. Dashed segments bridge unavailable rounds; they are not additional measurements. All curves compare clients, not successive global models.

J CHECKPOINT-SPECIFIC ACCURACY DIAGNOSTICS

The accuracy trajectories and private-branch interventions below use archived checkpoints. They provide checkpoint-specific diagnostics separate from the main benchmark tables. Evaluation rounds are shown explicitly.

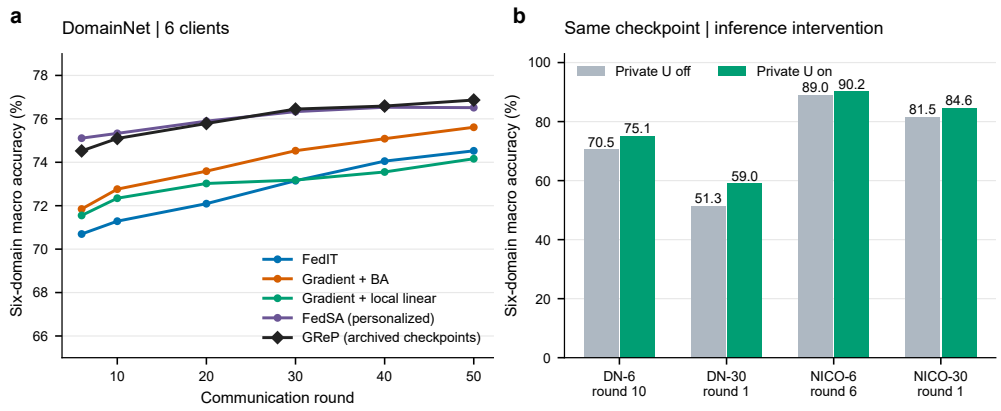


Figure 5: Measured accuracy trajectories and private-branch intervention. Left: DomainNet-6 control trajectories and archived GReP-LoRA checkpoints. Right: two evaluations of each identical checkpoint. The different settings are at different rounds; these bars do not rank the settings.

K WHY RESIDUAL GRADIENTS, AND WHAT DOES THE PRIVATE CORE LEARN?

This section distinguishes three statements: a residual is a local objective contrast; under additional curvature assumptions it identifies a personalized correction; and a trainable core fits that correction within fixed coordinates. None of these statements alone establishes improved held-out accuracy. Matrices can be vectorized whenever a Hessian is used below.

K.1 RESIDUAL GRADIENTS DIFFERENTIATE A CLIENT FROM THE POPULATION OBJECTIVE

At the same model W, define $F(W) = \sum_i p_i f_i(W)$ and $h_i(W) = f_i(W) - F(W)$. Then

$$R_i(W) = G_i(W) - \bar{G}(W) = \nabla h_i(W). \tag{23}$$

Thus the residual is not a heuristic subtraction without an objective: it is exactly the derivative of a client's loss relative to the population loss. This is an operational notion of client specificity, not a semantic classification of knowledge.

The relevant orthogonality is across clients. Equip tuples of client matrices with $\langle X, Y\rangle_p = \sum_i p_i \langle X_i, Y_i\rangle_F$ and define the consensus space $\mathcal{C} = \{(D, \dots, D) : D \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}\}$. The orthogonal projection of $G = (G_i)_i$ onto $\mathcal{C}$ is exactly $(\bar{G}, \dots, \bar{G})$, since

$$\sum_i p_i \|G_i - D\|_F^2 = \sum_i p_i \|R_i\|_F^2 + \left\|D - \bar{G}\right\|_F^2 .$$

Consequently $R = (R_i)_i \perp \mathcal{C}$. This is an exact decomposition in the product of the client and weight spaces. It does not imply $\langle R_i, \bar{G}\rangle_F = 0$ for an individual client. For example, if $G_i = a_i M$ for a common rank-one matrix M, all gradients and residuals share the same matrix direction, but $R_i = (a_i - \sum_j p_j a_j)M$ still describes genuine differences in the coefficients requested by clients. A method forcing every private matrix direction to be orthogonal to the shared matrix space would discard these differences. Separate per-client truncation and subsequent core training do not in general preserve the zero-mean tuple constraint.

Proposition 8 (Descent of the client contrast)**.** *Let h_i have an L_h-Lipschitz gradient on a region containing the update, and let $\mathcal{P}_i$ be any Frobenius-orthogonal projector onto a frozen private matrix space. For $D_i = -\eta \mathcal{P}_i R_i(W)$,*

$$h_i(W + D_i) \le h_i(W) - \eta \left(1 - \frac{L_h \eta}{2}\right) \|\mathcal{P}_i R_i(W)\|_F^2 . \tag{24}$$

In particular, $0 < \eta < 2/L_h$ decreases the contrast strictly when $\mathcal{P}_i R_i(W) \neq 0$.

Proof. Smoothness gives $h_i(W + D_i) \le h_i(W) + \langle R_i, D_i\rangle + (L_h/2)\|D_i\|_F^2$. Self-adjointness and idempotence give $\langle R_i, \mathcal{P}_i R_i\rangle = \|\mathcal{P}_i R_i\|_F^2$. Substitution proves the claim. □

Contrast descent does not imply descent of f_i on its own. The implemented core is trained on f_i, not on h_i. This proposition explains direction selection and is not a description of the implemented core update.

K.2 WHEN A RESIDUAL IDENTIFIES THE PERSONALIZED OPTIMUM

Consider an exact quadratic reference model,

$$f_i(w) = c_i + \tfrac{1}{2}(w - w_i^\star)^\top H (w - w_i^\star), \qquad 0 < \mu I \preceq H \preceq L I, \tag{25}$$

with a common Hessian across clients. Let $w_g^\star = \sum_i p_i w_i^\star$.

Theorem 9 (Residual-to-personalization correspondence)**.** *In Eq. (25), $w_g^\star$ is the unique minimizer of F, and at every common calibration point w,*

$$R_i(w) = H(w_g^\star - w_i^\star), \qquad w_i^\star - w_g^\star = -H^{-1} R_i(w). \tag{26}$$

The residual is independent of the calibration point in this reference model. If $H = \lambda I$, the optimal personal correction is $-R_i/\lambda$. For an orthogonal private projector $\mathcal{P}_i$, the best correction within its range at $w_g^\star$ is $D_i^\star = -\mathcal{P}_i R_i/\lambda$, and

$$f_i(w_g^\star) - \min_{D_i \in \mathrm{range}(\mathcal{P}_i)} f_i(w_g^\star + D_i) = \frac{\|\mathcal{P}_i R_i\|_F^2}{2\lambda}, \tag{27}$$

$$\min_{D_i \in \mathrm{range}(\mathcal{P}_i)} f_i(w_g^\star + D_i) - f_i(w_i^\star) = \frac{\|(I - \mathcal{P}_i) R_i\|_F^2}{2\lambda}. \tag{28}$$

For paired top-r residual singular bases, these expressions are respectively $\sum_{j \le r} \sigma_j(R_i)^2/(2\lambda)$ and $\sum_{j > r} \sigma_j(R_i)^2/(2\lambda)$.

Proof. $\nabla F(w) = H(w - \sum_i p_i w_i^\star)$, yielding the global optimum. Subtracting this derivative from $\nabla f_i(w) = H(w - w_i^\star)$ proves Eq. (26). In the isotropic case, complete the square:

$$f_i(w_g^\star + D) - f_i(w_g^\star) = \langle R_i, D \rangle + \frac{\lambda}{2} \|D\|_F^2 .$$

The component of R_i in the private space is optimized by $D = -\mathcal{P}_i R_i/\lambda$; its orthogonal complement cannot be changed. The last claim follows from Theorem 2. □

This proves a precise benefit of residual directions under a common, isotropic local model. For anisotropic H, the correction is curvature preconditioned; ordinary unweighted SVD of R_i need not be optimal for the personal correction. In a nonlinear network these identities are local approximations requiring evidence, not exact global properties.

A curvature counterexample. Let $f_1(w) = w^2/2$, $f_2(w) = w^2$, and $p_1 = p_2 = 1/2$. Both clients have the same optimum $w^\star = 0$, but at $w = 1$ the residuals are $-1/2$ and $1/2$. Nonzero residuals can reflect unequal curvature rather than different preferred optima. Sampling noise is another source. These counterexamples rule out the unconditional statement that every residual direction is personalized knowledge.

K.3 The core is a coefficient matrix in a tensor-product basis

Write the frozen bases as $B_i^p = \kappa Q_{L,i}$ and $A_i^p = \kappa Q_{R,i}^\top$, where both Q matrices have orthonormal columns. Set $\rho = s\kappa^2$. The effective private correction is

$$D_i(U_i) = \rho Q_{L,i} U_i Q_{R,i}^\top = \rho \sum_{j=1}^{r} \sum_{k=1}^{r} (U_i)_{jk} q_{L,ij} q_{R,ik}^\top. \tag{29}$$

The bases determine the available input and output subspaces; the r^2 entries of U_i determine coefficients of their outer products. Diagonal U_i changes only the r paired modes. A full core can also mix the $r(r-1)$ off-diagonal pairs. It is not constrained to be nonnegative, diagonal, orthogonal, or symmetric, and is therefore neither Σ nor a rotation matrix.

The full-core family depends on the two subspaces, not their particular orthonormal coordinates. For orthogonal $O_L, O_R \in \mathbb{R}^{r \times r}$,

$$(Q_L O_L)(O_L^\top U O_R)(Q_R O_R)^\top = Q_L U Q_R^\top.$$

This identity follows by canceling $O_L O_L^\top$ and $O_R O_R^\top$. An arbitrary coordinate change need not preserve diagonal cores. Moreover, exchanging cores between different client bases does not preserve the represented correction; branch-transfer tests must exchange both bases and the core together.

Proposition 10 (Core geometry and chain rule)**.** *For the parameterization in Eq.* (29)*,*

$$\|D_i(U_i)\|_F = \rho \|U_i\|_F , \qquad \sigma_j(D_i(U_i)) = \rho \sigma_j(U_i), \quad j \le r, \tag{30}$$

$$\nabla_{U_i} f_i = \rho Q_{L,i}^\top G_i Q_{R,i}, \qquad \Delta D_i = -\eta \rho^2 \mathcal{P}_i G_i \tag{31}$$

for one ordinary core-SGD step with the remaining parameters held fixed. At a common model, the core gradient decomposes exactly as

$$\nabla_{U_i} f_i = \rho Q_{L,i}^\top \bar{G} Q_{R,i} + \rho Q_{L,i}^\top R_i Q_{R,i}. \tag{32}$$

Proof. Left and right multiplication by matrices with orthonormal columns preserves the nonzero singular values and Frobenius norm. Apply the differential $dD_i = \rho Q_{L,i}(dU_i)Q_{R,i}^\top$ and the Frobenius inner product to obtain the gradient. Substituting a core gradient step into D_i gives the projected matrix update. The last identity is linearity and $G_i = \bar{G} + R_i$. □

When G_i is the gradient with respect to the effective weight at a deterministic layer, this is the ordinary matrix chain rule. With adapter input dropout, the same reasoning uses the gradient with respect to the adapter's linear correction acting on the dropped input; identifying it with the undropped whole-layer gradient requires dropout to be disabled. The actual forward uses a common dropout realization across adapter terms.

Eq. (32) makes a limitation explicit: residual initialization does not enforce purely residual learning throughout training. The core can also repair common directions not captured by the shared model. After clients acquire different private states, their gradients are evaluated at different models; a new common-model audit is needed before interpreting their difference as a data-only effect.

Why train the core instead of copying singular values? The calibration residual is only an initial, local loss derivative. Setting U_i to a fixed negative diagonal spectrum would choose one particular first-order step and step size. Training U_i lets the model fit amplitudes, signs and cross-mode couplings as the representation and shared factors change. At calibration, the residual-only projected gradient is diagonal for exact paired SVD bases. Off-diagonal gradients can arise from the mean term, anisotropic curvature or later changes in the network. Their usefulness is an empirical question.

K.4 Conditional convergence of the private core

Let $J_i = Q_{R,i} \otimes Q_{L,i}$, so $J_i^\top J_i = I$ and $\mathrm{vec}(D_i) = \rho J_i u_i$. Consider the local quadratic $q(d) = q(0) + g^\top d + d^\top H_i d/2$ with $\mu I \preceq H_i \preceq LI$, keeping the shared model and bases fixed. Then

$$\nabla^2_{u_i} q = \rho^2 J_i^\top H_i J_i, \qquad \mu\rho^2 I \preceq \nabla^2_{u_i} q \preceq L\rho^2 I. \tag{33}$$

Its unique minimizer and gradient-descent contraction are

$$u_i^\star = -\rho^{-1}(J_i^\top H_i J_i)^{-1} J_i^\top g, \tag{34}$$

$$q(\rho J_i u_i^k) - q(\rho J_i u_i^\star) \le \chi^{2k}[q(\rho J_i u_i^0) - q(\rho J_i u_i^\star)], \tag{35}$$

$$\chi = \max_{\lambda \in [\mu, L]} |1 - \eta\rho^2\lambda| < 1, \qquad 0 < \eta < 2/(L\rho^2). \tag{36}$$

To prove the bound, diagonalize $J_i^\top H_i J_i$ and apply the scalar gradient recursion to each eigencomponent. This is a convergence statement for a fixed quadratic core problem. It does not assert global convexity of a ViT or remove the drift terms in Theorem 6.

In the isotropic case, let G_i be the matrix with $\mathrm{vec}(G_i) = g$, and write $C_i = Q_{L,i}^\top G_i Q_{R,i}$. Optimizing over a full core instead of a diagonal core reduces the best quadratic loss by

$$q^\star_{\mathrm{diag}} - q^\star_{\mathrm{full}} = \frac{\|C_i - \mathrm{diag}(\mathrm{diag}(C_i))\|_F^2}{2\lambda}. \tag{37}$$

Here $\mathrm{diag}(\mathrm{diag}(C_i))$ denotes retaining only diagonal entries. The proof is coordinatewise completion of the square. For a pure calibration residual in its exact paired SVD bases, this quantity is zero. Consequently, a claimed benefit of full U must be demonstrated at later gradients or through independent training, not inferred merely from SVD.

A matched random-space reference. For independent uniformly random orthonormal $Q_L \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ and $Q_R \in \mathbb{R}^{d_{\mathrm{in}} \times r}$, and a fixed residual R,

$$\mathbb{E}\left\|Q_L^\top R Q_R\right\|_F^2 = \frac{r^2}{d_{\mathrm{out}} d_{\mathrm{in}}} \|R\|_F^2 .$$

To see this, write the squared norm as $\mathrm{tr}(R^\top Q_L Q_L^\top R Q_R Q_R^\top)$, use independence, and substitute $\mathbb{E}[Q_L Q_L^\top] = (r/d_{\mathrm{out}})I$ and $\mathbb{E}[Q_R Q_R^\top] = (r/d_{\mathrm{in}})I$. This distinguishes data-selected directions from merely adding r^2 trainable coefficients. It concerns a fixed residual and random bases, not a theorem that trained random adapters have a prescribed accuracy.

K.5 DISTINGUISHING STABLE RESIDUAL SIGNAL FROM SAMPLING NOISE

Suppose two disjoint, independently sampled calibration sets give unbiased $\widehat{G}_i^{(a)} = G_i + \varepsilon_i^{(a)}$, $a \in \{1, 2\}$. With fixed client weights, set $\widehat{R}_i^{(a)} = \widehat{G}_i^{(a)} - \sum_j p_j \widehat{G}_j^{(a)}$. Assume errors are zero mean and independent between the two sets. Then

$$\mathbb{E}\left\langle \widehat{R}_i^{(1)}, \widehat{R}_i^{(2)} \right\rangle = \|R_i\|_F^2 . \tag{38}$$

Indeed, expanding the inner product removes both linear noise terms and the cross-replicate noise term by zero mean and independence. For client-independent errors within one replicate, if $\mathbb{E}\,\|\varepsilon_j\|_F^2 = v_j$, then the residual noise energy is

$$\mathbb{E}\left\|\widehat{R}_i - R_i\right\|_F^2 = (1 - p_i)^2 v_i + \sum_{j \neq i} p_j^2 v_j . \tag{39}$$

Centering removes a common gradient, not estimation noise. Even the second moment of a held-out projected residual contains a noise floor; energy alone is insufficient evidence of signal.

Let $\widehat{\mathcal{P}}_i$ be estimated only from replicate 1. Conditional on that replicate,

$$\mathbb{E}_2\left\langle \widehat{\mathcal{P}}_i \widehat{R}_i^{(1)}, \widehat{R}_i^{(2)} \right\rangle = \left\langle \widehat{\mathcal{P}}_i \widehat{R}_i^{(1)}, R_i \right\rangle . \tag{40}$$

This is an out-of-sample directional prediction score, not a nonnegative quantity guaranteed by construction. The completed split-half experiment therefore reports held-out capture together with projected alignment, and compares own-client, foreign-client and random bases. For finite data sampled without replacement, independence is a population sampling idealization; the experiment is reported descriptively rather than treated as independent replicated trials.

L MEASURED EVIDENCE FOR RESIDUAL SELECTION AND PRIVATE-CORE MIXING

L.1 DO RESIDUAL DIRECTIONS PREDICT DISJOINT TRAINING DATA?

For each client, a deterministic shuffle selects 2,048 training examples, split into two disjoint sets of 1,024. At a fixed common pretrained model and common classifier initialization, we compute dense target-layer gradients on each set without updating any model parameter. Client sample-count weights define both mean gradients. Bases are selected using the first set only; the second set is never used for selection. No test labels enter this procedure. For layer ℓ, define $R_{i\ell}^{(a)}$ on split a and the projector $\mathcal{P}_{i\ell}$ selected on split 1. We report

$$\text{capture} = \frac{\sum_{i,\ell} \left\| \mathcal{P}_{i\ell} R_{i\ell}^{(2)} \right\|_F^2}{\sum_{i,\ell} \left\| R_{i\ell}^{(2)} \right\|_F^2}, \tag{41}$$

$$\text{prediction} = \frac{1}{150} \sum_{i,\ell} \frac{\left\langle \mathcal{P}_{i\ell} R_{i\ell}^{(1)}, R_{i\ell}^{(2)} \right\rangle}{\left\| \mathcal{P}_{i\ell} R_{i\ell}^{(1)} \right\|_F \left\| R_{i\ell}^{(2)} \right\|_F}. \tag{42}$$

The pooled sums apply no additional client weights. Foreign-basis scores average over all five other clients' bases per recipient (750 records). One common mean-gradient basis is used per layer for all clients.

Table 5: Selection on one training subset, evaluation on a disjoint subset. Capture is an energy ratio, not classification accuracy.

Private basis source	Held-out capture (%)	Prediction cosine
Own residual $G_i - \bar{G}$	57.21	0.689
Uncentered local G_i	43.49	0.533
Common mean $\bar{G}$	40.47	0.575
Other client's residual	32.00	0.512
Random orthonormal	0.030	0.011

Own residual bases give the largest capture and directional prediction in this diagnostic. All 150 own-residual projected cross-split inner products are positive. This supports repeatable client-dependent signal beyond selection-set fit; it does not establish noise-free residuals or a test-accuracy ordering after training. The finite-data sampling caveat in Appendix K applies.

Private spaces can overlap common directions. Using the original full calibration bank, own residual bases capture 97.07% of pooled calibration residual energy and 72.54% of pooled mean-gradient energy. Uncentered bases capture 66.90% and 79.11%, respectively; mean bases capture 60.64% and 98.84%. Thus residual selection does not produce a matrix space containing only private information. The calibration bank and split diagnostic have different sample budgets, so 97.07% versus 57.21% is not a controlled estimate of a generalization gap.

L.2 WHICH COMPONENTS OF THE LEARNED CORE AFFECT PREDICTIONS?

We hold the round-20 shared factors and classifier fixed and change only private branches at inference. A deterministic test subset selects at most eight examples per class and domain: 799 clipart, 773 infograph, 780 painting, 800 quickdraw, 800 real and 790 sketch images (4,742 total). Accuracy is the unweighted mean of six domain accuracies on this same subset. It is not the full-test benchmark score in the main tables.

The full own-client branch gives 74.26%, versus 71.15% with $U = 0$, 71.38% with only $\mathrm{diag}(U)$, 73.96% with only the off-diagonal entries, and 70.68% with U replaced by $\mathrm{tr}(U)I/r$. The off-diagonal share of pooled squared core norm increases from 68.59% at round 1 to 79.42%, 82.15% and 85.47% at rounds 6, 10 and 20. Together, these measurements show that the current fitted solution uses cross-mode coefficients extensively and functionally. They do not show that retraining a diagonal core must fail: alternative training can adapt the shared branch to the restriction.

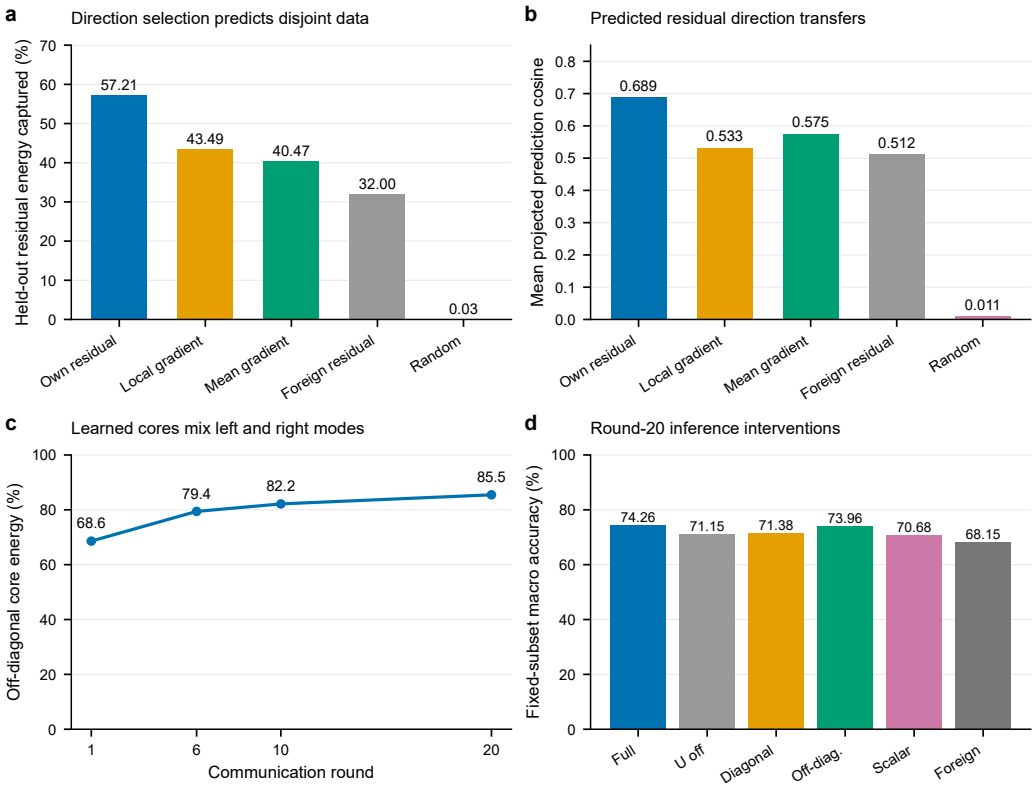


Figure 6: Residual selection and private-core diagnostics. (a,b) Disjoint training-subset capture and directional prediction from Table 5. (c) Off-diagonal squared-norm fraction pooled over clients and modules. (d) Round-20 fixed-checkpoint interventions on 4,742 test images; the vertical axis is truncated. "Foreign" averages whole-branch transfers over all five donors per recipient. No error bars are shown because this is one seed and one fixed data split.

L.3 Does a private branch preferentially serve its owner?

For each recipient, we evaluate all six complete private branches with the same shared model and test examples. A transferred branch includes the frozen left basis, learned core, and frozen right basis together. Each of the six recipients performs best with its own branch. Averaging foreign donors within each recipient and then recipients gives 68.15%, a 6.11 percentage-point deficit relative to own branches (74.26%). This supports recipient specificity of the trained branch. It does not isolate which part of that specificity originates from residual selection rather than local core optimization; the training controls below provide complementary evidence.

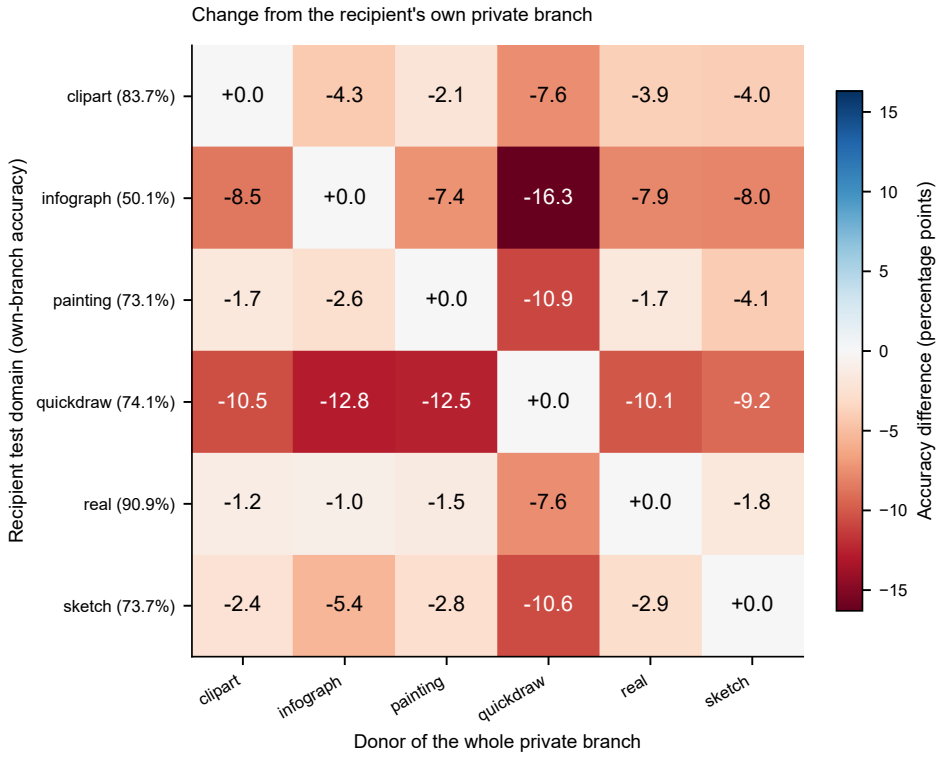


Figure 7: Whole-private-branch transfer at round 20. Each cell gives the recipient's accuracy change in percentage points relative to its own branch, using the same fixed test subset. Every off-diagonal entry is negative. Diagonal zeros are references, not an independent evaluation condition. Different clients' bare cores are never compared or exchanged without their associated bases.

L.4 MATCHED TRAINING CONTROLS AT A FIXED CHECKPOINT

Table 6 shows that the choice of private directions substantially affects personalized performance. Residual-selected bases with a full core achieve the highest accuracy, exceeding random bases and common mean-gradient bases by 2.83 and 1.17 percentage points, respectively. Disabling the learned private branch reduces accuracy by 4.57 points for residual bases, compared with only 0.10 points for random bases. A full core also outperforms a diagonal core by 1.83 points, although this comparison changes both expressivity and parameter count. These results support gradient-informed private-space selection and flexible adaptation within that space.

Table 6: Matched DomainNet-6 training at round 6, seed 51. "U off" disables the private branch at the same checkpoint; it is not an independently trained no-private model.

Private basis / core	Personalized (%)	U off (%)
Own residual / full core	74.52	69.95
Common mean gradient / full core	73.35	70.26
Random orthonormal / full core	71.69	71.59
Own residual / diagonal core	72.69	71.00

M FROM FACTOR INCREMENTS TO INTRINSIC UPDATE GEOMETRY

This appendix connects the implemented local-linear model to geometric aggregation. It distinguishes an ambient gradient, a finite matrix increment, and a tangent vector. The last two need not coincide. We omit the fixed LoRA scale s below; restoring it scales effective corrections accordingly.

M.1 THE REPRESENTED INCREMENT AND THE FACTORIZATION ARE DIFFERENT OBJECTS

For a server adapter $X = BA$ and a locally trained complete-BA adapter,

$$Z_i = (B + \delta B_i)(A + \delta A_i) - BA = \underbrace{B\delta A_i + \delta B_i A}_{\xi_i} + \delta B_i \delta A_i. \tag{43}$$

Z_i is determined by the two endpoint matrices. Changing either endpoint's factorization leaves it unchanged. In contrast, $\delta A_i, \delta B_i$ require compatible factor coordinates. The implemented local shared branch directly uses $X + \xi_i$, so its fitted increment is ξ_i; it is not obtained by training the full product and subsequently deleting the cross term.

At a common model, a dense gradient G_i is also independent of adapter factorization. It specifies an infinitesimal Euclidean descent direction, not the finite increment generated by multiple steps of factor SGD. Consequently, our residual initialization is a first-order choice of adaptation space, not an estimate of every later local matrix increment.

M.2 THREE ORTHOGONAL TANGENT COMPONENTS

Assume $\mathrm{rank}(X) = r$ and write its compact SVD as $X = L\Sigma V^\top$. Let $P_L = LL^\top$ and $P_V = VV^\top$. The embedded fixed-rank geometry (Vandereycken, 2013) gives

$$T_X\mathcal{M}_r = \{LMV^\top + KV^\top + LN^\top : L^\top K = 0,\ V^\top N = 0\}, \tag{44}$$

$$\Pi_X Z = P_L Z P_V + (I - P_L) Z P_V + P_L Z (I - P_V). \tag{45}$$

These three terms respectively describe coefficient changes within the existing input/output spaces, output-space variations, and input-space variations. Their pairwise Frobenius inner products vanish because $P_L(I - P_L) = 0$ and $P_V(I - P_V) = 0$. The remaining block $(I - P_L)Z(I - P_V)$ is normal to the tangent space. These labels describe matrix geometry; none assigns a global or personalized role to a component.

To connect Eq. (45) to our local branch, observe that $(I - P_L)B = 0$ and $A(I - P_V) = 0$. Hence $(I - P_L)\xi_i(I - P_V) = 0$ and $\Pi_X \xi_i = \xi_i$. Conversely, in coordinates $B = L\Sigma$, $A = V^\top$, any tangent vector $LMV^\top + KV^\top + LN^\top$ is represented by $\delta B = LM + K$ and $\delta A = \Sigma^{-1}N^\top$. Thus $B\delta A + \delta B A$ spans the tangent space at full rank. The factor pair is redundant: adding $(BH, -HA)$ to $(\delta B, \delta A)$ changes neither ξ nor its three matrix components, consistent with quotient representations (Mishra et al., 2014).

What this does and does not imply. The available shared directions have a tangent interpretation. Ordinary factor SGD nevertheless induces $G \mapsto BB^\top G + GA^\top A$, rather than the orthogonal projector Π_X; see Appendix E. Under $(B, A) \mapsto (BR, R^{-1}A)$ this operator generally changes. Selecting directions from G_i is factorization independent, but the implemented optimizer is not invariant to arbitrary factor rescaling.

M.3 Intrinsic aggregation requires an explicit matrix-space update

An intrinsic tangent-aggregation construction would first compute Z_i relative to the same server matrix, form $\zeta_i = \Pi_X Z_i$, combine those tangent vectors, and retract the result. For an unweighted geometry, one example is

$$\eta = \sum_i p_i \zeta_i, \qquad X^+ = \mathcal{R}_X(\eta) = \mathrm{SVD}_r(X + \eta). \tag{46}$$

Locally around a rank-r reference, truncated SVD supplies a smooth retraction; its domain requires the relevant singular-value separation (Vandereycken, 2013; Boumal, 2023). Projection and subsequent rank restoration have a long history in low-rank optimization and dynamics (Koch & Lubich, 2007). The resulting rule is independent of factor choices for fixed endpoint matrices. Gauge-dependent client optimization can still produce different endpoint matrices, so this is an aggregation property, not automatically an end-to-end training invariance.

The implemented update instead gives, with $\bar{\delta}A = \sum_i p_i \delta A_i$ and $\bar{\delta}B = \sum_i p_i \delta B_i$,

$$X^+ = (B + \bar{\delta}B)(A + \bar{\delta}A) = X + \sum_i p_i \xi_i + \bar{\delta}B\bar{\delta}A. \tag{47}$$

It preserves a rank-at-most-r shared factorization, but it is not Eq. (46). For full local BA training, the discrepancy from mean client products is the matrix covariance term

$$\sum_i p_i \delta B_i \delta A_i - \bar{\delta}B\bar{\delta}A = \sum_i p_i (\delta B_i - \bar{\delta}B)(\delta A_i - \bar{\delta}A).$$

This is the aggregation issue addressed by product-preserving methods such as FedEx-LoRA (Singhal et al., 2025b), not a new identity. Rewriting factor averaging using deltas alone does not remove it.

Anisotropic metrics need a specified action. Denote the three tangent-block projectors in Eq. (45) by Π_c. A positive block metric $M_X = \sum_{c=1}^{3} a_c \Pi_c$, $a_c > 0$, is independent of the choice of orthonormal bases within P_L, P_V. Blockwise means and dispersions of $\Pi_c \zeta_i$ could inform a_c. However, for one common metric,

$$\arg\min_{\eta \in T_X \mathcal{M}_r} \sum_i p_i \left\| \eta - \zeta_i \right\|_{M_X}^2 = \sum_i p_i \zeta_i,$$

because differentiation gives $M_X(\eta - \sum_i p_i \zeta_i) = 0$. Therefore this barycenter alone provides no adaptive aggregation. An explicit operation such as $M_X^{-1} \sum_i p_i \zeta_i$ would change component weights; its descent and convergence properties require analysis of that operation and its estimated coefficients. Neither such a metric nor its retraction is implemented in the reported personalized runs.

Rank and initialization qualifications. If one initial factor is zero, $X = 0 \notin \mathcal{M}_r$ for $r > 0$; a smooth rank-r tangent model requires a nondegenerate reference or a separately specified rank-deficient treatment. The gradient initializer uses two nonzero factors, with a compensating frozen backbone. The private correction $B_i^p U_i A_i^p$ lies in a fixed linear matrix space and need not be tangent to the shared adapter. Its zero initial core is well defined there. The combined shared/private adapter has rank at most $r_g + r_p$, not a single rank-r_g budget.

Scope of the geometric comparison. The present method implements residual-based allocation of shared and private adaptation. Common tangent projection, adaptive block metrics, and SVD retraction specify a different server rule. The geometric comparison clarifies the meaning of the local first-order model; none of these alternative aggregation operations is part of GReP-LoRA.

N Detailed analysis: from similar factors to sharing decisions

This section supplies the additional proofs behind the revised main text. The statements are organized by their logical dependencies: a counterexample to a sharing statistic, a criterion for selecting spaces, the value of those spaces, and the behavior of the actual gradient operator. Unless stated otherwise, a matrix denotes one target layer with all other coordinates held fixed. Several target layers and a shared head can be included by direct sums; layer-dependent scales must then be retained. Exact zero residuals with zero bases contribute the zero projector, rather than an orthonormal basis.

The inverse below is taken on independent basis coordinates; for an empty selected space, the improvement is defined as zero.

N.1 Perfect factor agreement with opposite effective updates

Proposition 11 (Factor agreement does not certify update agreement)**.** *There exist two strongly convex quadratic losses in effective weight coordinates and a common nonzero rank-one factor initialization such that one simultaneous ordinary factor-SGD step produces cosine similarity one for both pairs of complete factors, but cosine similarity minus one for the two nonzero effective weight increments. The same construction has opposite first-order increments under the local-linear forward. This existence statement uses nonzero initial A and B; it is not a statement about the first step of conventional zero-B initialization.*

Proof. Let $a \in \mathbb{R}^{d_{\text{in}}}$ and $b \in \mathbb{R}^{d_{\text{out}}}$ be unit vectors, and $M = ba^\top$. Set $s = 1$, $B_0 = b$, $A_0 = a^\top$ and $W_f = 0$. For $q > 0$, define two client objectives on effective weights,

$$f_\pm(X) = \tfrac{1}{2} \|X - (1 \pm q)M\|_F^2 . \tag{48}$$

Both are nonnegative, smooth, strongly convex matrix objectives. At the common initial matrix $X_0 = M$, their gradients are $G_\pm = \mp qM$. The ordinary factor chain rule gives

$$\nabla_A f_\pm = B_0^\top G_\pm = \mp q a^\top, \tag{49}$$

$$\nabla_B f_\pm = G_\pm A_0^\top = \mp q b. \tag{50}$$

With $\epsilon = \eta q \in (0, 1)$, one simultaneous plain-SGD step therefore gives

$$A_\pm = (1 \pm \epsilon)a^\top, \qquad B_\pm = (1 \pm \epsilon)b. \tag{51}$$

All four scaling coefficients are positive, so both corresponding-factor cosines equal one exactly. However,

$$Z_\pm = B_\pm A_\pm - B_0 A_0 = (\pm 2\epsilon + \epsilon^2)M. \tag{52}$$

Since $2\epsilon + \epsilon^2 > 0$ and $-2\epsilon + \epsilon^2 < 0$, the normalized Frobenius inner product of the two nonzero increments is -1. Their individual factor increments also have cosine -1. In the local-linear forward, the effective increments are $\pm 2\epsilon M$, so the same sign disagreement persists. No client gauge change, different anchor, scale asymmetry between A and B, or nonconvex weight loss is required. □

What follows, and what does not. Even the largest possible value of complete-factor cosine provides no nontrivial universal lower bound on effective-update cosine. The example is a logical obstruction to inferring update agreement from that statistic. It does not estimate how frequently this happens in real training. The empirical reversal, cumulative-versus-current comparison, and scale controls address the latter question separately. Cosine measures direction only; neither the example nor a high empirical cosine asserts equal magnitudes.

N.2 Why mean and residual are the relevant orthogonal split

Define $\mathcal{C} = \{(D, \dots, D)\}$ in the weighted client product space. For any common D,

$$\sum_i p_i \|G_i - D\|_F^2 = \sum_i p_i \left\|R_i + (\bar{G} - D)\right\|_F^2 \tag{53}$$

$$= \sum_i p_i \|R_i\|_F^2 + \left\|\bar{G} - D\right\|_F^2 . \tag{54}$$

The cross term vanishes because $\sum_i p_i R_i = 0$. Thus the unique closest consensus tuple is $(\bar{G})_i$, and R is the orthogonal residual in *client product space*. Moreover, $R_i = \nabla(f_i - F)$ follows by differentiation. This gives an exact optimization meaning to a client contrast, without labeling an individual singular vector as semantically global or private.

The decomposition is also the solution of a simple local decision problem. For a common correction d, the population quadratic upper model $\langle \bar{G}, d \rangle + L \|d\|^2 / 2$ is minimized by $-\bar{G}/L$. Allowing separate corrections instead gives $-G_i/L$ for client i. Their departures from the common choice are $-R_i/L$, and the reduction in this surrogate objective is $\sum_i p_i \|R_i\|^2 / (2L)$. This argument compares the stated quadratic models. Upper bounds on two different losses cannot by themselves prove the same ordering of actual losses; Theorem 12 makes a direct comparison under explicit quadratic assumptions.

N.3 Conditional objective value of a fixed private space

Theorem 12 (Value of a fixed private space)**.** *Let $W_\star$ satisfy $\nabla_W F(W_\star) = 0$ in effective weight coordinates, with all remaining model parameters fixed. Suppose*

$$f_i(W_\star + D) - f_i(W_\star) = r_i^\top d + \tfrac{1}{2} d^\top H_i d, \qquad r_i = \mathrm{vec}(R_i(W_\star)), \quad d = \mathrm{vec}(D),$$

where $\mu I \preceq H_i \preceq LI$ and $0 < \mu \leq L$. For any fixed orthonormal private bases, let $J_i = Q_{R,i} \otimes Q_{L,i}$ and keep $\mathcal{S}_i$ as in Eq. (10)*. Its best attainable improvement is*

$$I_i = f_i(W_\star) - \min_{D \in \mathcal{S}_i} f_i(W_\star + D) = \tfrac{1}{2} r_i^\top J_i (J_i^\top H_i J_i)^{-1} J_i^\top r_i,$$

$$\frac{\|\mathcal{P}_i R_i(W_\star)\|_F^2}{2L} \leq I_i \leq \frac{\|\mathcal{P}_i R_i(W_\star)\|_F^2}{2\mu}. \tag{55}$$

For the zero space, set $I_i = 0$ without forming an inverse. If, additionally, the bases are exact paired leading singular vectors of $R_i(W_\star)$ itself, then $I_i \geq (\mu/L) I^\star_{i,r_p}$, where $I^\star_{i,r_p}$ is the best improvement over all rank-at-most-r_p corrections. The two-sided bound permits frozen bases selected at a different calibration point; the last rank-competitive statement requires bases selected at the reference point appearing in the bound.

Proof. At the shared stationary point, $\bar{G} = 0$, hence the client gradient equals $R_i = R_i(W_\star)$. Let $r = \mathrm{vec}(R_i)$, $J = Q_R \otimes Q_L$, and $H = H_i$. Orthonormality gives $J^\top J = I$. The range of the scaled map ρJ is the same as that of J for $\rho \neq 0$, so use effective coefficients c. The restricted quadratic is

$$q(c) = r^\top J c + \tfrac{1}{2} c^\top H_S c, \qquad H_S = J^\top H J. \tag{56}$$

For any c, $\mu \|c\|^2 \leq c^\top H_S c \leq L \|c\|^2$. Thus H_S is positive definite and its unique minimizer is $c^\star = -H_S^{-1} J^\top r$. Completing the square gives

$$q(c) = \tfrac{1}{2}(c - c^\star)^\top H_S (c - c^\star) - \tfrac{1}{2} r^\top J H_S^{-1} J^\top r. \tag{57}$$

The second term is minus the attainable improvement I_i. The eigenvalues of H_S^{-1} lie in $[1/L, 1/\mu]$, and $\left\|J^\top r\right\|^2 = \|\mathcal{P}_i R_i\|_F^2$. This proves the two-sided bound in Eq. (55).

For the rank-constrained comparison, let $S_r = \sum_{j \leq r_p} \sigma_j(R_i)^2$. Any rank-at-most-$r_p$ correction D satisfies

$$\langle R_i, D \rangle + \tfrac{1}{2} \mathrm{vec}(D)^\top H \, \mathrm{vec}(D) \geq \langle R_i, D \rangle + \tfrac{\mu}{2} \|D\|_F^2 . \tag{58}$$

Completing the square and applying the rank approximation result from Theorem 2 shows that the minimum of the right side over rank-at-most-r_p matrices is $-S_r/(2\mu)$. Consequently $I^\star_{i,r_p} \leq S_r/(2\mu)$. The selected residual space attains at least $S_r/(2L)$, so

$$I_i \geq \frac{S_r}{2L} \geq \frac{\mu}{L} I^\star_{i,r_p}. \tag{59}$$

This remains valid when $S_r = 0$, with both sides zero. The ratio depends on curvature conditioning; exact optimality follows for $H = \lambda I$, but is not asserted for arbitrary anisotropic H. □

Nonstationary references and the meaning of stationarity. At an arbitrary quadratic reference W, the same proof applies with $r_i = \mathrm{vec}(\nabla_W f_i(W))$. The corresponding bounds therefore use $\mathcal{P}_i \nabla_W f_i(W)$, not only $\mathcal{P}_i R_i(W)$. Furthermore, vanishing shared factor gradients imposes $B^\top \nabla_W F = 0$ and $(\nabla_W F)A^\top = 0$; these conditions can hold with $\nabla_W F \neq 0$. Thus factor stationarity does not supply the full effective-weight stationarity assumed in the theorem.

Corollary 13 (Calibration-to-reference bridge)**.** *Under Theorem 12, keep $\mathcal{P}_i$ fixed from calibration at W_0, and define $\varepsilon_i^{\mathrm{drift}} = \|R_i(W_\star) - R_i(W_0)\|_F$. Then Eq.* (14) *holds for this space.*

Proof. An orthogonal projector is contractive, so the reverse triangle inequality gives

$$\|\mathcal{P}_i R_i(W_\star)\|_F \geq \left[\|\mathcal{P}_i R_i(W_0)\|_F - \varepsilon_i^{\mathrm{drift}}\right]_+ .$$

Square this inequality and apply the lower bound in Theorem 12. For the zero projector both sides vanish. No stationarity assumption is imposed on W_0. □

A local nonlinear extension. For a nonlinear loss, let $w_\star = \mathrm{vec}(W_\star)$, $H_i = \nabla^2 f_i(w_\star)$, and suppose $\mu I \preceq H_i \preceq LI$. Here I_i denotes the attainable gain in the quadratic Taylor model, and $d^\star$ its restricted minimizer. If the Hessian is M_H-Lipschitz on the segment from $w_\star$ to $w_\star + d^\star$, Taylor's theorem gives

$$f_i(w_\star) - f_i(w_\star + d^\star) \geq I_i - \tfrac{M_H}{6}\|d^\star\|^3, \qquad \|d^\star\| \leq \frac{\left\|J^\top r\right\|}{\mu}. \tag{60}$$

This transfers a sufficiently small local-model gain to the true loss with an explicit remainder. It does not establish a global quadratic model for ViT or ensure that every training step lies in the required neighborhood.

N.4 Calibration away from a common optimum: curvature mismatch

Consider now exact client quadratics with different Hessians,

$$f_i(w) = c_i + \tfrac{1}{2}(w - w_i^\star)^\top H_i (w - w_i^\star), \quad H_i \succeq \mu I, \quad \bar{H} = \sum_i p_i H_i. \tag{61}$$

The population optimum is $w_g^\star = \bar{H}^{-1} \sum_i p_i H_i w_i^\star$, which need not equal the weighted average of the client optima. At a common calibration point w_0, put $v = w_0 - w_g^\star$. Direct expansion gives

$$G_i(w_0) = H_i v + H_i(w_g^\star - w_i^\star), \tag{62}$$

$$\bar{G}(w_0) = \bar{H} v, \tag{63}$$

$$R_i(w_0) = H_i(w_g^\star - w_i^\star) + (H_i - \bar{H})v. \tag{64}$$

The middle equality uses $\sum_i p_i H_i (w_g^\star - w_i^\star) = 0$. Rearranging the last equation yields

$$w_i^\star - w_g^\star = -H_i^{-1} R_i(w_0) + H_i^{-1}(H_i - \bar{H})v, \tag{65}$$

$$\left\|w_i^\star - w_g^\star + H_i^{-1} R_i(w_0)\right\| \leq \frac{\left\|H_i - \bar{H}\right\|_2}{\mu} \|v\| . \tag{66}$$

Thus residuals identify curvature-preconditioned personal displacements exactly when curvatures agree, or at the population optimum; away from it, heterogeneous curvature introduces a quantified error.

If $\|H_i - \lambda I\|_2 \leq \delta_i$, comparison to the unpreconditioned residual gives

$$\left\|w_i^\star - w_g^\star + R_i(w_0)/\lambda\right\| \leq \frac{\delta_i}{\lambda\mu} \|R_i(w_0)\| + \frac{\left\|H_i - \bar{H}\right\|_2}{\mu} \|v\| . \tag{67}$$

To verify the first coefficient, use $\lambda^{-1} I - H_i^{-1} = \lambda^{-1} H_i^{-1}(H_i - \lambda I)$. The common-curvature identity in Theorem 9 is therefore an exact special case, not an assumption hidden inside a general claim.

N.5 Sampling error and the lifetime of a frozen residual basis

Suppose each $G_i(w)$ is L_i-Lipschitz on a common-model region, and let $\bar{L} = \sum_j p_j L_j$, $\ell_i = L_i + \bar{L}$. Then

$$\|R_i(w) - R_i(w_0)\|_F \le \ell_i \|w - w_0\|. \tag{68}$$

This follows from the triangle inequality applied to $G_i - \sum_j p_j G_j$. Let $\widehat{R}_i$ estimate $R_i(w_0)$ with error at most ε_i. For a fixed orthogonal projector $\mathcal{P}_i$ chosen at calibration,

$$\|\mathcal{P}_i R_i(w)\|_F \ge \big(\left\|\mathcal{P}_i \widehat{R}_i\right\|_F - \varepsilon_i - \ell_i \|w - w_0\| \big)_+, \tag{69}$$

$$\|(I - \mathcal{P}_i) R_i(w)\|_F \le \tau_i + 2\varepsilon_i + \delta_i^{\mathrm{svd}} + \ell_i \|w - w_0\|, \quad \tau_i^2 = \sum_{j > r_p} \sigma_j(R_i(w_0))^2. \tag{70}$$

The first line uses reverse triangle inequality and projector contraction. For the second, apply projector contraction to Eq. (68) and the approximate-SVD bound proved earlier; δ_i^{svd} denotes the additive approximation error in residual norm, not a singular-vector distance. At a stationary reference satisfying Theorem 12, the squared positive part in the first line, divided by $2L$, is a valid lower bound on restricted improvement.

The inequalities identify three distinct limitations: spectral truncation, calibration estimation/computation, and subsequent movement. They do not guarantee useful capture for arbitrarily long training. Residual drift is defined at common weights; subtracting gradients of already personalized models also mixes in model-state differences. The held-out diagnostics in Appendix L address sampling stability, not a proof that the frozen spaces remain optimal throughout all rounds.

N.6 Why a full core can help, and why SVD alone does not prove it

In fixed orthonormal bases write $C = Q_L^\top G Q_R$ and $D(U) = \rho Q_L U Q_R^\top$. For an isotropic quadratic,

$$q(U) - q(0) = \rho \langle C, U \rangle + \tfrac{\lambda \rho^2}{2} \|U\|_F^2. \tag{71}$$

Full-core and diagonal-core minima have $U^\star_{\mathrm{full}} = -C/(\lambda\rho)$ and $U^\star_{\mathrm{diag}} = -\operatorname{diag}(\operatorname{diag} C)/(\lambda\rho)$. Subtracting the completed squares proves Eq. (37). Ordinary core GD from zero uses the scalar contraction $a = 1 - \eta\lambda\rho^2$ in each allowed coordinate. For $0 < \eta < 2/(\lambda\rho^2)$, the finite-step difference is exactly

$$q(U^k_{\mathrm{diag}}) - q(U^k_{\mathrm{full}}) = \frac{1 - a^{2k}}{2\lambda} \|\operatorname{offdiag} C\|_F^2. \tag{72}$$

At initial paired residual bases and a residual-only objective, C is diagonal. Nonzero off-diagonal benefit must arise from a different current gradient, common-gradient leakage, or non-isotropic coupling. This motivates the diagonal retraining control, rather than declaring full-core superiority from parameter count alone. A full core also stays within the original left and right subspaces; it cannot represent missing ambient directions.

N.7 The actual one-step factor/core update

Fix shared anchors A, B and current client cores, and disable dropout, regularization, clipping and momentum for this one deterministic step. Client i may have its own effective starting weight W_i through its core; $G_i = \nabla f_i(W_i)$ and $\bar{G} = \sum_i p_i G_i$ in this subsection. Unlike calibration, these need not be gradients at one common weight. The comparison freezes the private core in one branch and trains it in the other, starting from identical states. Gradients and shared increments at that start are therefore identical between the two branches.

Set $\rho = s\kappa^2$ and define operators on client tuples by

$$\begin{aligned} &K_g(X) = s^2(BB^\top X + X A^\top A), \\ &(\mathcal{K}_0 G)_i = K_g(\bar{G}), \qquad (\mathcal{K}_p G)_i = \rho^2 \mathcal{P}_i G_i, \qquad \mathcal{K}_+ = \mathcal{K}_0 + \mathcal{K}_p. \end{aligned} \tag{73}$$

For a zero private branch, use $\mathcal{P}_i = 0$.

Proposition 14 (Actual step and additional descent)**.** *Under the deterministic, full-participation, same-start conditions above, one simultaneous factor/core SGD step of size η changes the evaluated weights by*

$$D_+ = -\eta \mathcal{K}_+ G + (E)_i, \qquad E = s\overline{\delta B}\,\overline{\delta A} = O(\eta^2).$$

The shared-only comparison has $D_0 = -\eta \mathcal{K}_0 G + (E)_i$ with the same E. If $\mathcal{F}(W_1, \dots, W_N) = \sum_i p_i f_i(W_i)$ is L_W-smooth on a region containing both update segments in the weighted product norm, then, with $h_G = \rho^2 \sum_i p_i \left\|\mathcal{P}_i G_i\right\|_F^2$,

$$\mathcal{F}(W + D_0) - \mathcal{F}(W + D_+) \geq \eta h_G - \frac{L_W}{2}\left(\|D_0\|_p^2 + \|D_+\|_p^2\right). \tag{74}$$

If $h_G > 0$, the difference is strictly positive for sufficiently small $\eta > 0$. This compares one step from an identical state with other model coordinates held fixed; it does not order independently trained trajectories or remove the server reanchoring term.

Proof. At zero local shared increments the actual local-linear forward gives

$$\delta A_i = -\eta s B^\top G_i, \quad \delta B_i = -\eta s G_i A^\top, \quad \delta U_i = -\eta\rho Q_{L,i}^\top G_i Q_{R,i}. \tag{75}$$

Full-participation averaging gives $\overline{\delta A} = -\eta s B^\top \bar{G}$ and $\overline{\delta B} = -\eta s \bar{G} A^\top$. Expanding the evaluated product, including the actual private-core change, proves

$$D_i = s(B\overline{\delta A} + \overline{\delta B} A) + \rho Q_{L,i} \delta U_i Q_{R,i}^\top + E \tag{76}$$

$$= -\eta K_g(\bar{G}) - \eta\rho^2 \mathcal{P}_i G_i + E, \tag{77}$$

$$E = \eta^2 s^3 \bar{G} A^\top B^\top \bar{G}, \qquad \|E\|_F \leq \eta^2 s^3 \|A\|_2 \|B\|_2 \left\|\bar{G}\right\|_F^2. \tag{78}$$

Thus the remainder is genuinely second order at fixed anchors and gradients; it is not deleted from the evaluated model.

For tuples X, Y, self-adjointness of K_g implies $\langle X, \mathcal{K}_0 Y\rangle_p = \langle \bar{X}, K_g \bar{Y}\rangle$, symmetric in X, Y. Moreover,

$$\langle X, \mathcal{K}_0 X\rangle_p = s^2(\left\|B^\top \bar{X}\right\|_F^2 + \left\|\bar{X} A^\top\right\|_F^2) \geq 0, \tag{79}$$

$$\langle X, \mathcal{K}_p X\rangle_p = \rho^2 \sum_i p_i \left\|\mathcal{P}_i X_i\right\|_F^2 \geq 0. \tag{80}$$

Both operators are therefore self-adjoint positive semidefinite. By Jensen and spectral submultiplicativity,

$$\|\mathcal{K}_+\|_{\mathrm{op}} \leq s^2(\|B\|_2^2 + \|A\|_2^2) + \rho^2 =: \Lambda. \tag{81}$$

For the linear endpoint $\widetilde{D} = -\eta \mathcal{K}_+ G$ and an L_W-smooth product loss, the descent lemma and $\|\mathcal{K}_+ G\|_p^2 \leq \Lambda \langle G, \mathcal{K}_+ G\rangle_p$ give

$$\mathcal{F}(W + \widetilde{D}) \leq \mathcal{F}(W) - \eta(1 - L_W \eta \Lambda/2) \langle G, \mathcal{K}_+ G\rangle_p. \tag{82}$$

The actual endpoint also contains $(E)_i$, bounded above. To compare actual shared-only and shared-plus-core endpoints, use the two-sided Taylor bound

$$|\mathcal{F}(W + D) - \mathcal{F}(W) - \langle G, D\rangle_p| \leq (L_W/2) \|D\|_p^2. \tag{83}$$

Subtracting the upper bound for D_+ from the lower bound for D_0 leaves

$$\langle G, D_0 - D_+\rangle_p = \eta \langle G, \mathcal{K}_p G\rangle_p = \eta h_G, \tag{84}$$

because the two E terms are equal. This proves Eq. (74). □

An explicit sufficient step-size condition. Let $a_j = \|\mathcal{K}_j G\|_p$ for $j \in \{0, +\}$, and let $c_E = s^3 \|A\|_2 \|B\|_2 \left\|\bar{G}\right\|_F^2$. For $0 < \eta \leq 1$, $\|D_j\|_p \leq \eta(a_j + c_E)$. If $h_G > 0$, any

$$0 < \eta < \min\left\{1, \frac{2h_G}{L_W[(a_0 + c_E)^2 + (a_+ + c_E)^2]}\right\} \tag{85}$$

with the required smoothness neighborhood guarantees strict improvement. This bound is sufficient and can be conservative. It compares one step from the same state; different trained trajectories can reach different gradients and do not inherit a pointwise ordering from this statement.

Relating actual core descent to residual capture. At a common model, define $h_R = \sum_i p_i \|\mathcal{P}_i R_i\|^2$ and $h_M = \sum_i p_i \|\mathcal{P}_i \bar{G}\|^2$. The reverse triangle inequality in product space gives

$$h_G/\rho^2 = \sum_i p_i \|\mathcal{P}_i(\bar{G} + R_i)\|^2 \geq (\sqrt{h_R} - \sqrt{h_M})_+^2. \tag{86}$$

Large residual capture alone need not imply a large actual step: the projected mean and residual can cancel. At a shared stationary point, $h_M = 0$ and the identity becomes $h_G = \rho^2 h_R$. The distinction is essential because the implemented core minimizes its local loss, not the contrast $f_i - F$ at every training step.

N.8 Momentum, local step counts and participation in the convergence bound

This subsection makes the direction error in Theorem 6 explicit. It does not assume the errors vanish in the recorded runs. Let $g_i = \nabla f_i(z_{i,0})$ be the round-start gradient in local trainable coordinates $z_i = (\theta, u_i)$. Coupled weight decay may be included in f_i if included consistently in both the endpoint objective and local surrogate.

Exact momentum unrolling. For reset momentum with $\beta \in [0, 1)$, no Nesterov term, and $v_{i,0} = 0$,

$$v_{i,k+1} = \beta v_{i,k} + h_{i,k}, \qquad z_{i,k+1} = z_{i,k} - \eta v_{i,k+1}, \tag{87}$$

where $h_{i,k}$ is the vector actually supplied to the momentum update. Induction gives $v_{i,k+1} = \sum_{j=0}^{k} \beta^{k-j} h_{i,j}$. Interchanging the two finite sums yields

$$z_{i,K_i} - z_{i,0} = -\eta \sum_{j=0}^{K_i - 1} c_{K_i,j} h_{i,j}, \tag{88}$$

$$c_{K,j} = \frac{1 - \beta^{K-j}}{1 - \beta}, \qquad C_K = \sum_{j=0}^{K-1} c_{K,j} = \frac{K}{1-\beta} - \frac{\beta(1-\beta^K)}{(1-\beta)^2}. \tag{89}$$

Thus a round with frozen gradient has effective step ηC_K, not $K\eta$. Unequal local epoch data sizes give different K_i and different effective steps even when the nominal learning rate is identical.

Drift and clipping. Assume $\|h_{i,j}\| \leq M$, local surrogate gradients are $\widetilde{L}$-Lipschitz, and their gradients match g_i at the anchor. Then $\|v_{i,j}\| \leq M/(1-\beta)$ and $\|z_{i,j} - z_{i,0}\| \leq \eta M j/(1-\beta)$. Let the conditional mean difference between $h_{i,j}$ and the local surrogate gradient at the current state have norm at most $b_{\mathrm{clip,i}}$. This can include clipping bias and any explicitly specified objective mismatch. For the normalized local direction $\widehat{d}_i = C_{K_i}^{-1} \sum_j c_{K_i,j} h_{i,j}$, conditional on selection,

$$\left\| \mathbb{E}_t \widehat{d}_i - g_i \right\| \leq \underbrace{\frac{\widetilde{L}\eta M}{1-\beta} \frac{\sum_j j c_{K_i,j}}{C_{K_i}} + b_{\mathrm{clip,i}}}_{e_i^{\max}}. \tag{90}$$

The proof is triangle inequality, conditional expectation, and the path bound. For task-gradient norm clipping $\mathrm{clip}_c(h)$ before adding coupled decay, the clipping part is at most $\mathbb{E}_t(\|h\| - c)_+$, since $\|\mathrm{clip}_c(h) - h\| = (\|h\| - c)_+$. Clipping an unbiased gradient is generally biased. A deterministic bound on $h_{i,j}$ additionally requires a bounded parameter region if coupled decay is included outside clipping.

Explicit coefficient error from selection and unequal steps. Fix a positive reference round step α and set $a_i = \eta C_{K_i}/\alpha$. For random selected set $\mathcal{S}$, define

$$\omega_i = \mathbb{E}_t[\mathbf{1}_{i \in \mathcal{S}} \tilde{p}_i a_i], \qquad \nu_i = \Pr_t(i \in \mathcal{S}) a_i. \tag{91}$$

Assume the round's K_i are fixed conditional on its initial state and Eq. (90) holds for every selection containing i. The conditional mean of the normalized shared direction is $\sum_i \omega_i g_{\theta,i}$ plus a drift term;

the corresponding private direction is $\nu_i g_{u,i}$ plus drift. Since the weighted objective gradient is $(\sum_i p_i g_{\theta,i}, g_{u,1}, \ldots, g_{u,m})$, if each initial local gradient has norm at most M_g then

$$\|b_{\theta,t}\| \le M_g \sum_i |\omega_i - p_i| + \sum_i \omega_i e_i^{\max}, \tag{92}$$

$$\|b_{u_i,t}\| \le M_g |\nu_i - 1| + \nu_i e_i^{\max}, \tag{93}$$

$$\|b_t\|_P^2 \le \left(M_g \|\omega - p\|_1 + \sum_i \omega_i e_i^{\max} \right)^2 + \sum_i p_i \left(M_g |\nu_i - 1| + \nu_i e_i^{\max} \right)^2. \tag{94}$$

To prove this, add and subtract the initial local gradients inside each weighted update, then bound their coefficient mismatch and local drift separately. Equation (94) can be substituted directly into Eq. (21); conditional update noise is defined by subtracting the conditional mean and must have the stated bounded variance.

For full participation and equal K, choose $\alpha = \eta C_K$. Then $\omega_i = p_i$ and $\nu_i = 1$. Without clipping or objective mismatch, Eq. (90) is $O(\alpha)$ for fixed K, β, yielding the same $O(T^{-1/2})$ stationarity rate with horizon-dependent $\alpha = O(T^{-1/2})$, including reset momentum. By contrast, partial participation, unequal steps and clipping need not yield vanishing errors under this norm. Some selection effects can be modeled using a different metric, but that requires restating the theorem; they are not silently dropped here. Fixed-rate runs receive an error-floor guarantee when the assumptions and bounds hold.