FEDRESPRO: BREAKING RESOURCE-INVERSE DISTRIBUTION DILEMMA VIA FEDERATED RESIDUAL PROJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneity has long been a fundamental obstacle to effective model collaboration in federated learning (FL). Despite extensive efforts to address data, system, and model heterogeneity, existing methods remain inadequate for a prevalent yet largely overlooked scenario, which we term Resource Inverse Distribution (RID), where data-rich frontline clients are constrained by limited model capacity, while high-capacity clients have access to only limited data. RID disperses statistical reliability and optimization capacity across different clients, inducing Statistical-Optimization Capability Decoupling: low-capacity clients cannot sufficiently explore effective directions beyond their constrained optimization spaces, whereas complementary knowledge discovered by high-capacity clients suffers from statistical uncertainty due to limited observations. To address this dilemma, we propose FedRespro, a federated heterogeneous model collaboration framework that recouples statistical reliability and optimization capacity across heterogeneous clients. Specifically, FedRespro extracts complementary residual knowledge from the expanded optimization spaces of high-capacity clients and suppresses unreliable information through statistical, spatial, and temporal reliability estimation. It further constructs rank-matched references via capacity-aware low-rank projection, enabling data-rich low-capacity clients to absorb complementary knowledge beyond their local exploration capabilities. By compensating for the statistical unreliability of high-capacity clients and the constrained optimization capability of data-rich clients, FedRespro enables effective cross-client complementarity between statistical and optimization advantages under RID scenarios. Experiments show an average server accuracy gain of 2.19 percentage points across three Qwen2 scales over the strongest evaluated baselines. On BERT-base, CoLA MCC ($\times 100$) improves by 23.43 points, while DBpedia accuracy improves by 16.81 percentage points, each relative to its strongest evaluated baseline. A reference implementation is available at `https://anonymous.4open.science/r/FedRespro/`.

1 INTRODUCTION

As AI models continue to scale, increasingly valuable data remain distributed across end devices and private data silos, motivating the widespread application of federated learning (FL) as a paradigm for collaborative training without directly sharing raw data Bai et al. (2024); Cheng et al. (2025). However, differences in data distributions, computational resources, and model capacities across clients substantially complicate effective collaboration Yi et al. (2024a); Wang et al. (2024a). Existing heterogeneous FL studies have extensively investigated statistical, system, model, and communication heterogeneity, as well as their coexistence in practical systems Chen & Vikalo (2024); Liu et al. (2024); Yi et al. (2024b). However, these factors are typically modeled as independent resource variations, overlooking structured dependencies across resource dimensions, where data availability and computational capacity may be systematically coupled rather than independently distributed.

In contrast to prior work, in this paper, we focus on a representative but underexplored configuration, termed **Resource Inverse Distribution (RID)**, where data availability and computational capacity are inversely distributed across clients. Specifically, frontline clients close to data-generating

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.

sources typically possess abundant local data but limited model capacity, whereas higher-tier computing nodes can support larger models but have less direct access to frontline data. Such resource asymmetry naturally arises in hierarchical distributed systems, including industrial IoT and mobile or wearable healthcare Wong et al. (2026); Jia et al. (2024); Xiang et al. (2024).

More importantly, RID goes beyond physical resource mismatch and induces a structural decoupling between statistical reliability and optimization capacity across clients, which we term *Statistical-Optimization Capability Decoupling (SOCD)*. Specifically, data-rich but low-capacity clients provide more reliable statistical estimates but are confined to restricted feasible optimization spaces, whereas high-capacity clients can explore a broader range of optimization directions but lack sufficient statistical support to reliably validate these additional directions due to limited local data. Consequently, under RID, neither type of client alone can simultaneously provide reliable statistical evidence and sufficient optimization capacity, making cross-capacity collaboration essential for complementing their respective strengths. As illustrated in Fig.1, this capability decoupling gives rise to two key challenges that distinguish RID from conventional FL heterogeneity: **Reliable Complementary Knowledge Discovery** and **Capacity-Constrained Knowledge Absorption**.

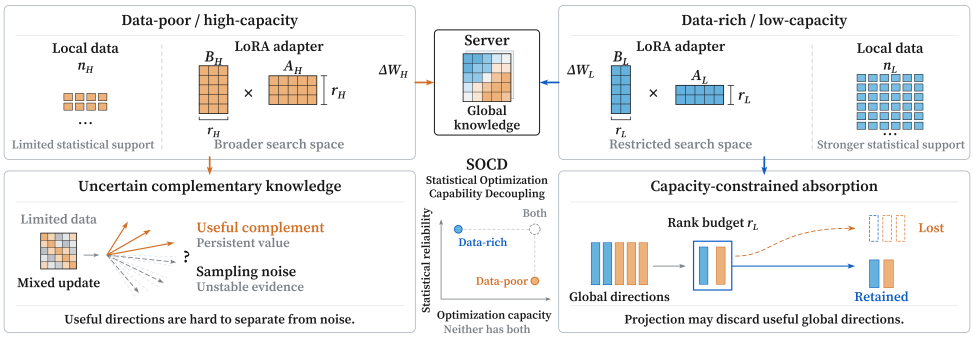

Figure 1: **The main challenges for heterogeneous model collaboration under RID.** Resource Inverse Distribution (RID) decouples statistical reliability from optimization capacity across heterogeneous clients, resulting in Statistical-Optimization Capability Decoupling (SOCD). Consequently, SOCD raises two key challenges: identifying reliable complementary directions from the broader but uncertain optimization spaces of high-capacity clients (*Reliable Complementary Knowledge Discovery*), and enabling capacity-limited clients to absorb high-dimensional complementary knowledge (*Capacity-Constrained Knowledge Absorption*).

First, RID prevents data-rich clients from fully exploiting their statistical advantages. Despite access to abundant training data, their limited computational capacity confines optimization to restricted feasible spaces, leaving statistically supported optimization information insufficiently explored and thereby creating a *Local Learnability Bottleneck*. The broader optimization spaces of high-capacity clients provide an opportunity to explore such missing information beyond the capacity of data-rich clients. However, due to limited local observations, the additional information discovered by high-capacity clients is not inherently reliable and may contain both valuable complementary knowledge and statistical perturbations. Therefore, the first challenge, which we term **Reliable Complementary Knowledge Discovery**, is to identify complementary information with persistent optimization value from limited and potentially unstable observations.

Second, reliable complementary knowledge acquired through cross-capacity collaboration cannot be directly absorbed by resource-constrained clients. Such knowledge may span a higher-dimensional optimization space whose complexity exceeds the representation capacity of constrained models. Consequently, valuable information already preserved through global collaboration may be lost when projected back to low-capacity clients, creating a *Knowledge Absorption Bottleneck*. This gives rise to the second challenge, **Capacity-Constrained Knowledge Absorption**: how to transform high-dimensional complementary knowledge into optimization information that constrained models can effectively absorb without increasing their computational or representational capacity.

To address these challenges, we propose FedRespro, a heterogeneous federated collaboration framework that recouples the statistical and optimization advantages separated by RID. To enable reliable complementary knowledge discovery, FedRespro accumulates cross-round residuals and evaluates their reliability against a statistical reference constructed from sample-weighted updates, jointly

considering parameter-space consistency and cross-round persistence to maintain a reliable global residual. To enable capacity-constrained knowledge absorption, FedRespro decouples global knowledge preservation from client-side transfer: the server retains the complete residual state, while rank-matched residual sketches transform it into capacity-compatible references for heterogeneous clients. By combining the broader optimization spaces of high-capacity clients with the statistical reliability of data-rich clients, FedRespro effectively mitigates the SOCD problem induced by RID. Three-seed experiments show server accuracy gains of 0.85, 4.79, and 0.93 percentage points over the strongest evaluated baselines on Qwen2-0.5B/1.5B/7B, respectively. Client-view gains depend on model scale, and the encoder configurations yield larger improvements as detailed in Section 4.1. Our main contributions are summarized as follows:

- We identify **Resource Inverse Distribution (RID)**, an underexplored heterogeneous FL setting that decouples statistical reliability from optimization capacity across clients.
- We propose **FedRespro**, which reliably discovers and preserves complementary knowledge through residual-based cross-client collaboration, and transfers it via capacity-aware rank-matched projection.
- Experiments across three Qwen2 scales show an average server accuracy gain of 2.19 percentage points over the strongest evaluated baselines. On BERT-base, CoLA MCC ($\times 100$) improves by 23.43 points, while DBpedia accuracy improves by 16.81 percentage points, each relative to its strongest evaluated baseline. Ablations quantify residual-memory effects, and client-view analyses identify scale-dependent deployment differences.

2 Related Works

2.1 Heterogeneous Federated Learning

Heterogeneous federated learning aims to enable effective collaboration among clients with diverse computational resources and model capacities. A representative line of work constructs resource-adaptive submodels from a shared global model. HeteroFL (Diao et al., 2021) trains width-scaled local models according to client capabilities, while FedRolex (Alam et al., 2022) employs rolling submodel extraction to improve parameter coverage across heterogeneous clients. ScaleFL (Ilhan et al., 2023) further adapts both network width and depth, and FedP3 (Yi et al., 2024a) formulates personalized pruning for client-specific resource constraints. FedLMT (Liu et al., 2024) instead exploits low-rank factorization to reduce local computation while providing convergence guarantees.

Another line of research enables knowledge exchange among structurally different models. pFedHR (Wang et al., 2023) dynamically reassembles heterogeneous models for personalized collaboration, whereas DFRD (Luo et al., 2023) transfers knowledge through data-free distillation. FedType (Wang et al., 2024a) introduces homogeneous proxy models and asymmetric reciprocity learning to bridge heterogeneous private models. FedMRL (Yi et al., 2024b) combines heterogeneous local models with a shared lightweight model through Matryoshka representations, while FedFree (Du et al., 2025) performs fine-grained layer-wise alignment without relying on public proxy data. Despite substantial progress, these methods primarily adapt collaboration mechanisms to differences in client resources or model structures. Data availability and model capacity are generally treated as independent or coexisting sources of heterogeneity, leaving their structured inverse coupling—and the resulting separation between statistical reliability and optimization capability—largely unexplored.

2.2 LoRA for Heterogeneous Model Collaboration

Low-Rank Adaptation (LoRA) has recently become an important mechanism for reducing the computation and communication costs of federated fine-tuning. Early work mainly focuses on improving the stability and correctness of federated LoRA aggregation. FFA-LoRA (Sun et al., 2024) freezes one LoRA factor to alleviate the mismatch between local joint optimization and server-side separate aggregation. FLoRA (Wang et al., 2024b) identifies aggregation errors in heterogeneous LoRA and introduces stacking-based aggregation to support different client ranks. FlexLoRA (Bai et al., 2024) further synthesizes client updates into a full-size LoRA representation and redistributes rank-adaptive adapters through SVD, explicitly exploiting heterogeneous client resources. FRLoRA (Yan

et al., 2025) accumulates residual low-rank updates in a higher-rank global parameter space and reinitializes local adapters to mitigate client drift.

More recent studies improve the expressivity and personalization of federated low-rank collaboration. RoLoRA (Chen et al., 2025) alternates the optimization of the two LoRA factors to improve robustness and expressiveness. Ravan (Raje et al., 2025) uses multiple low-rank heads to recover richer update spaces under constrained communication, while FedLEASE (Wang et al., 2025) adaptively allocates and selects LoRA experts according to client representations. FLoRG (Meng et al., 2026) reformulates federated low-rank aggregation through Gram matrices and Procrustes alignment to reduce aggregation and decomposition drift. These methods substantially advance federated LoRA under data, task, and rank heterogeneity, but do not explicitly consider the inverse coupling between data availability and model capacity. In contrast, FedRespro focuses on determining whether additional directions explored by high-capacity but data-scarce clients are statistically reliable, and on transferring the resulting complementary knowledge to clients according to their heterogeneous representational capacities.

3 Proposed Method

In this section, we present **FedRespro**, a federated heterogeneous model collaboration framework for RID scenarios. FedRespro discovers reliable complementary knowledge from heterogeneous client residuals, preserves it through cross-round residual accumulation, and projects the resulting global knowledge into rank-matched references for clients with different model capacities. An overview of FedRespro is illustrated in Fig. 2.

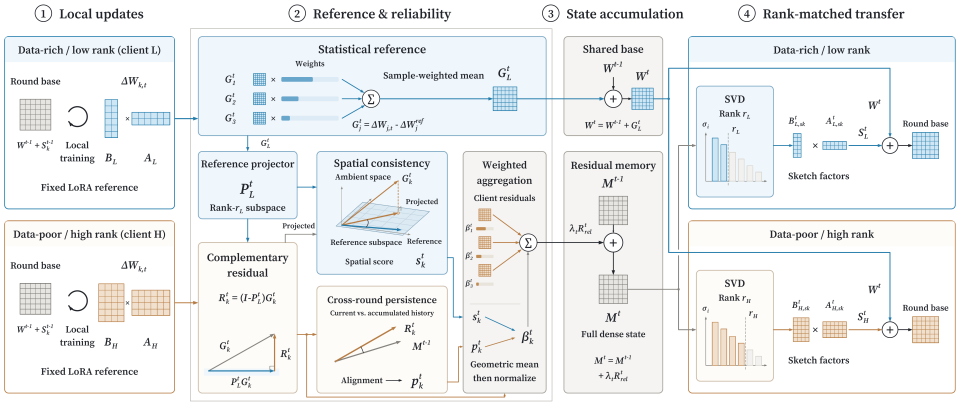

Figure 2: **Overview of FedRespro.** FedRespro evaluates client residuals using the statistical reference (G^t), parameter-space consistency (S^t), and cross-round persistence (P^t), and accumulates reliable complementary knowledge into a server-side residual. The server then performs rank-matched projection to generate capacity-adaptive references for subsequent local training for heterogeneous clients under RID scenarios.

3.1 Reliable Complementary Knowledge Discovery

Under RID, data-rich frontline clients are confined to limited optimization spaces, while high-capacity clients can explore complementary directions but lack sufficient statistical support to validate them reliably. FedRespro therefore characterizes newly acquired knowledge through client update residuals relative to the reference model, which are subsequently assessed and aggregated according to their reliability.

Client residual construction. Let $\mathcal{K}$ denote the participating clients, partitioned into relatively data-rich low-capacity clients $\mathcal{L}$ and high-capacity clients $\mathcal{H}$. Each client k caches a fixed adaptive rank reference $\Delta\mathbf{W}_k^{\text{ref}}$, which remains unchanged across communication rounds. After local training at round t, its update residual is defined as

$$\mathbf{G}_k^t = \Delta\mathbf{W}_{k,t} - \Delta\mathbf{W}_k^{\text{ref}}. \tag{1}$$

Considering the statistical reliability alone does not fully characterize the reliability of complementary knowledge provided by high-capacity models. Instead of directly aggregating all residuals according to clients' local data volume, FedRespro further evaluates complementary knowledge from two additional dimensions: *spatial consistency*, which measures its alignment with statistically supported optimization directions, and *cross-round persistence*, which captures its temporal stability across communication rounds.

Statistical reference. Since data-rich clients provide stronger statistical support, FedRespro constructs a sample-size-weighted reference

$$\mathbf{G}_{\mathcal{L}}^{t} = \sum_{j \in \mathcal{L}} \frac{n_j}{N_{\mathcal{L}}} \mathbf{G}_{j}^{t}, \qquad N_{\mathcal{L}} = \sum_{j \in \mathcal{L}} n_j. \tag{2}$$

This reference serves as a statistical anchor for evaluating the additional directions explored by high-capacity clients.

Let $\mathcal{P}_{\mathcal{L}}^{t}(\cdot)$ denote the rank-$r_{\mathcal{L}}$ projection defined by the truncated SVD of the statistical reference. For $k \in \mathcal{H}$, we decompose

$$\mathbf{G}_{k}^{t} = \underbrace{\mathcal{P}_{\mathcal{L}}^{t}(\mathbf{G}_{k}^{t})}_{\mathbf{G}_{k,\parallel}^{t}} + \underbrace{(\mathbf{I} - \mathcal{P}_{\mathcal{L}}^{t})\mathbf{G}_{k}^{t}}_{\mathbf{R}_{k}^{t}}, \tag{3}$$

where $\mathbf{G}_{k,\parallel}^{t}$ captures the shared component and $\mathbf{R}_{k}^{t}$ represents the complementary residual beyond the statistically supported subspace.

Spatial consistency. Given the inherent statistical uncertainty of high-capacity clients, their updates should remain compatible with the statistically supported optimization tendency within the shared subspace. We measure this shared-space agreement by

$$s_{k}^{t} = \left[\frac{\left\langle \mathcal{P}_{\mathcal{L}}^{t}(\mathbf{G}_{k}^{t}), \bar{\mathbf{G}}_{\mathcal{L}}^{t} \right\rangle_{F}}{\|\mathcal{P}_{\mathcal{L}}^{t}(\mathbf{G}_{k}^{t})\|_{F} \left\|\bar{\mathbf{G}}_{\mathcal{L}}^{t}\right\|_{F} + \epsilon} \right]_{+}, \tag{4}$$

where $\bar{\mathbf{G}}_{\mathcal{L}}^{t} = \mathcal{P}_{\mathcal{L}}^{t}(\mathbf{G}_{\mathcal{L}}^{t})$, $\langle \cdot, \cdot \rangle_{F}$ and $\| \cdot \|_{F}$ denote the Frobenius inner product and norm, respectively, and $[x]_{+} = \max(x, 0)$. A larger s_{k}^{t} indicates stronger agreement between client k and the statistically supported optimization tendency in their shared space. More importantly, after normalization, $2(1 - s_{k}^{t})$ corresponds to the squared directional disagreement observed in this shared subspace. Under the nested reference-aligned setting specified in Appendix A, the rank-$r_{\mathcal{L}}$ reference space is contained in the rank-r_k space accessible to high-capacity client k. Their effective dimensions are therefore $r_{\mathcal{L}}^2$ and r_k^2, leaving $r_k^2 - r_{\mathcal{L}}^2$ complementary dimensions. This yields the following reliability estimator.

Theorem 1 (Reliability Estimation from Spatial Consistency) *Let $r_{\mathcal{L}}$ denote the effective rank of the statistically supported low-capacity subspace and $r_k > r_{\mathcal{L}}$ the effective rank accessible to high-capacity client k. For $s_{k}^{t} > 0$, the statistical uncertainty of its complementary directions can be estimated by*

$$\widehat{\mathcal{U}}_{k,\perp}^{t} = \frac{r_k^2 - r_{\mathcal{L}}^2}{r_{\mathcal{L}}^2} \, 2(1 - s_{k}^{t}). \tag{5}$$

Proof sketch. For Frobenius-normalized projected updates, cosine similarity gives the exact shared-space disagreement $\|\hat{\mathbf{G}}_{k}^{t} - \hat{\mathbf{G}}_{\mathcal{L}}^{t}\|_{F}^{2} = 2(1 - s_{k}^{t})$. Since the rank-$r_{\mathcal{L}}$ shared space and the additional space of client k contain $r_{\mathcal{L}}^2$ and $r_k^2 - r_{\mathcal{L}}^2$ effective dimensions, respectively, the bounded-noise assumption allows this disagreement to be extrapolated by their dimension ratio, yielding Eq. 5. See Appendix A.4 for the complete derivation.

Theorem 1 links the observable consistency in the shared low-rank space to the statistical uncertainty of the unobserved complementary directions. Furthermore, we provide an approximation error bound for this reliability estimation in **Corollary 1**.

Corollary 1 (Approximation Error Bound for Complementary Reliability Estimation) *Under the conditions of Theorem 1, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\left| \widehat{\mathcal{U}}_{k,\perp}^{t} - \mathcal{U}_{k,\perp}^{t} \right| \leq \Delta_{k}^{t}(\delta), \tag{6}$$

where

$$\Delta_k^t(\delta) = C_{\mathrm{rel}}\sqrt{\log\frac{6}{\delta}}\left(N_{\mathcal{L}}^{-1/2} + r_{\mathcal{L}}^{-1} + (r_k^2 - r_{\mathcal{L}}^2)^{-1/2}\right). \tag{7}$$

Here, C_{rel} is a finite problem-dependent constant determined by the reference-estimation, subspace-concentration, and bounded-anisotropy constants specified in Appendix A.

Proof sketch. The three terms arise from finite-sample estimation of the statistical reference, concentration over the shared reference-aligned subspace, and concentration over the complementary subspace, respectively. Their high-probability deviations are combined by a union bound. The complete derivation is provided in Appendix A.4.

Cross-round persistence. Considering spatial consistency alone cannot distinguish transient perturbations from complementary directions that persist across communication rounds. We evaluate the temporal consistency of $\mathbf{R}_k^t$ as

$$p_k^t = \begin{cases} 1, & t = 1 \text{ or } \|\mathbf{M}^{t-1}\|_F \le \epsilon_M, \\ \left[\dfrac{\langle \mathbf{R}_k^t, \mathbf{M}^{t-1}\rangle_F}{\|\mathbf{R}_k^t\|_F \|\mathbf{M}^{t-1}\|_F + \epsilon}\right]_+, & \text{otherwise.} \end{cases} \tag{8}$$

The evaluation procedure initializes the residual memory from the first reliable complementary observation, after which p_k^t favors directions that remain consistent with previously preserved knowledge.

Reliability-aware residual aggregation. Although s_k^t is only an observable proxy for complementary reliability, **Corollary 1** shows that its induced reliability estimate admits a bounded approximation error. FedRespro therefore directly combines spatial consistency with cross-round persistence. Let

$$Z_t = \sum_{h\in\mathcal{H}} \sqrt{s_h^t p_h^t}. \tag{9}$$

The reliability weight is defined as

$$\beta_k^t = \begin{cases} \dfrac{\sqrt{s_k^t p_k^t}}{Z_t}, & Z_t > 0, \\ 0, & Z_t = 0, \end{cases} \qquad k \in \mathcal{H}. \tag{10}$$

Accordingly, the reliable complementary residual is

$$\mathbf{R}_{\mathrm{rel}}^t = \sum_{k\in\mathcal{H}} \beta_k^t \mathbf{R}_k^t. \tag{11}$$

When $Z_t = 0$, no reliable complementary evidence is available and $\mathbf{R}_{\mathrm{rel}}^t = \mathbf{0}$, leaving the complementary memory unchanged in that round. Finally, the statistically supported update and the reliable complementary residual jointly form

$$\mathbf{G}^t = \mathbf{G}_{\mathcal{L}}^t + \lambda_t \mathbf{R}_{\mathrm{rel}}^t, \tag{12}$$

where λ_t is a hyperparameter that controls the contribution of complementary knowledge. $\mathbf{G}_{\mathcal{L}}^t$ updates the shared base state, $\lambda_t \mathbf{R}_{\mathrm{rel}}^t$ updates the complementary residual state.

3.2 Capacity-Aware Rank-Matched Residual Projection

FedRespro preserves complementary knowledge in a dense server-side state $\mathbf{M}^t = \mathbf{M}^{t-1} + \lambda_t \mathbf{R}_{\mathrm{rel}}^t$, while $\mathbf{W}^t$ maintains the statistically supported shared state. To adapt the preserved knowledge to heterogeneous client capacities, given $\mathbf{M}^t = \mathbf{U}^t \mathbf{\Sigma}^t (\mathbf{V}^t)^\top$, FedRespro constructs the rank-matched residual reference

$$\mathbf{S}_k^t = \mathbf{U}_{1:r_k}^t \mathbf{\Sigma}_{1:r_k}^t (\mathbf{V}_{1:r_k}^t)^\top. \tag{13}$$

By the Eckart–Young–Mirsky theorem, $\mathbf{S}_k^t$ is the optimal rank-r_k approximation of $\mathbf{M}^t$ under the Frobenius norm.

Rather than transmitting $\mathbf{S}_k^t$ densely, FedRespro encodes it as rank-r_k Residual-Sketch factors with $\mathcal{O}(r_k(d_{\text{out}} + d_{\text{in}}))$ scalars. The sketch is incorporated into the subsequent round-specific optimization base, while the fixed reference $\Delta\mathbf{W}_k^{\text{ref}}$ in Eq. 1 remains unchanged. Thus, FedRespro decouples dense server-side knowledge preservation from capacity-matched low-rank utilization without expanding client model capacity.

3.3 Theoretical Analysis

Convergence guarantee. **Theorem 2** shows that FedRespro preserves the standard stochastic optimization convergence behavior, while its additional error is explicitly controlled by reliable complementary knowledge discovery and capacity-aware rank-matched projection. With $\eta = \Theta(T^{-1/2})$, the convergence rate becomes $\mathcal{O}(T^{-1/2}) + \mathcal{O}(T^{-1}\sum_{t=0}^{T-1} B_t^2)$, and the complete proof is provided in Appendix B. For theoretical analysis, $\mathbf{W}^t$ denotes the effective server state $\mathbf{W}_{\text{base}}^t + \mathbf{M}^t$.

Theorem 2 (Convergence of FedRespro) *Let $F^\star$ denote the lower bound of the global objective. Under the standard smoothness, bounded stochastic variance, and bounded client-heterogeneity assumptions detailed in Appendix B, for a learning rate $\eta \le 1/(4L)$, FedRespro satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(\mathbf{W}^t)\right\|_F^2 \le \underbrace{\frac{2\left(F(\mathbf{W}^0) - F^\star\right)}{\eta T}}_{\textit{Initial optimality gap}} + \underbrace{L\eta\sigma^2}_{\textit{Stochastic variance}} + \underbrace{\frac{5}{2T}\sum_{t=0}^{T-1} B_t^2}_{\textit{FedRespro approximation error}}, \tag{14}$$

where

$$B_t^2 = 3L^2\varepsilon_{\text{mem},t}^2 + 3L^2\varepsilon_{\text{rank},t}^2 + 3\varepsilon_{\text{loc}}^2, \tag{15}$$

with $\varepsilon_{\text{mem},t}$, $\varepsilon_{\text{rank},t}$, and ε_{loc} denoting the accumulated reliability-memory error, rank-projection error, and local-update error, respectively.

Computational overhead. FedRespro introduces additional computation mainly at the server, while leaving the asymptotic client-side training cost unchanged. For an update matrix with $D = d_{\text{out}}d_{\text{in}}$, the dominant server-side overhead is

$$\mathcal{O}(D(r_{\mathcal{L}} + r_{\max}))\,, \tag{16}$$

which remains moderate since $r_{\mathcal{L}}, r_{\max} \ll \min(d_{\text{out}}, d_{\text{in}})$. Detailed analysis is provided in Appendix C.1.

Communication overhead. FedRespro preserves the accumulated dense residual only at the server side and transmits a rank-s_k Residual-Sketch to client k, requiring only

$$\mathcal{O}(s_k(d_{\text{out}} + d_{\text{in}})) \tag{17}$$

additional scalars instead of $\mathcal{O}(d_{\text{out}}d_{\text{in}})$ for dense residual transmission. For this separate-sketch protocol with $s_k = r_k$, the total communication is approximately $1.5\times$ that of standard low-rank federated training, while avoiding dense residual transmission. Thus, FedRespro decouples dense server-side knowledge preservation from capacity-matched low-rank communication. Details in Appendix C.2. The adapter-only protocol measured in Section 4.3 is reported separately and does not include this additional residual channel.

4 Experiments

Datasets and models. We fine-tune Qwen2-0.5B/1.5B/7B on Commonsense170K and report mean accuracy across eight reasoning tasks (500 fixed examples each). We also evaluate BERT-base-uncased on CoLA (8,551 training and 1,043 validation examples; MCC) and DBpedia-14 (30,000

training and 3,000 test examples; accuracy). **Federated settings.** Qwen uses 40 clients, six per round, 40 rounds, and ranks 4/8 in 75/25% proportions. BERT uses 20 clients, five per round, ranks 2/4/8, and 20/15 rounds for CoLA/DBpedia. Smaller ranks are assigned to data-rich clients to instantiate RID. Both suites use AdamW and query/value adapters; BERT also trains its classifier. Tables report three training seeds for every Qwen main-table entry. We compare HetLoRA (Cho et al., 2024), FlexLoRA (Bai et al., 2024), FLoRA (Wang et al., 2024b), RB-LoRA (Ha et al., 2026), and FSLoRA (Fang et al., 2025). Full experimental protocols are given in Appendix D.

4.1 Main Results

Table 1: Eight-task mean accuracy (%) after 40 training rounds. Values are means over three seeds, with sample SD as subscripts. R4/R8 denote client deployment views and S the server; bold marks the best mean.

	Qwen2-0.5B			Qwen2-1.5B			Qwen2-7B		
Method	R4	R8	S	R4	R8	S	R4	R8	S
HetLoRA	$37.71_{2.20}$	$38.56_{1.37}$	$38.98_{1.48}$	$56.84_{0.59}$	$56.53_{0.30}$	$56.59_{0.23}$	$77.08_{0.11}$	$77.09_{0.22}$	$77.04_{0.04}$
FlexLoRA	$39.69_{1.04}$	$39.25_{1.38}$	$39.52_{1.73}$	$58.56_{0.60}$	$58.42_{0.78}$	$58.44_{0.81}$	$77.96_{0.14}$	$78.03_{0.24}$	$77.93_{0.13}$
FLoRA	$39.80_{0.54}$	$39.80_{0.54}$	$39.76_{0.88}$	$54.96_{0.82}$	$54.96_{0.82}$	$54.97_{0.44}$	$74.79_{0.27}$	$74.79_{0.27}$	$74.80_{0.20}$
RB-LoRA	$38.91_{2.07}$	$37.50_{1.78}$	$37.62_{1.84}$	$58.42_{0.54}$	$58.41_{0.71}$	$58.39_{0.48}$	$78.34_{0.04}$	$78.34_{0.14}$	$78.20_{0.18}$
FSLoRA	$\mathbf{40.28}_{0.81}$	$\mathbf{40.40}_{0.89}$	$40.39_{0.81}$	$58.24_{1.10}$	$57.95_{0.55}$	$58.03_{0.48}$	$78.62_{0.45}$	$78.92_{0.43}$	$78.91_{0.43}$
FedRespro	$40.02_{3.53}$	$39.14_{3.84}$	$\mathbf{41.24}_{3.63}$	$\mathbf{61.74}_{1.29}$	$\mathbf{61.50}_{1.21}$	$\mathbf{63.23}_{1.18}$	$\mathbf{79.13}_{0.16}$	$\mathbf{79.51}_{0.18}$	$\mathbf{79.84}_{0.07}$

Qwen. FedRespro achieves the best mean accuracy in seven of the nine columns of Table 1, leading the server comparison at all three scales. Its server accuracies of 41.24%, 63.23%, and 79.84% exceed the strongest baselines by 0.85, 4.79, and 0.93 percentage points. The largest advantage appears at 1.5B, where R4 and R8 also improve by 3.18 and 3.08 points over their strongest baselines. At 7B, FedRespro leads in both client views and achieves a server standard deviation of 0.07, compared with FSLoRA's 0.43, combining higher accuracy with lower variation across seeds.

Table 2: BERT-base-uncased results on CoLA (MCC $\times 100$) and DBpedia (accuracy, %). Three-seed means use sample SD subscripts. Bold marks the best mean. "failed" denotes failure to converge.

Dataset	HetLoRA	FlexLoRA	FLoRA	RB-LoRA	FSLoRA	Fed-Respro
CoLA	failed	$6.17_{1.4}$	$0.60_{1.0}$	$19.57_{11.3}$	$11.55_{7.1}$	$\mathbf{43.00}_{0.8}$
DBpedia	$15.58_{0.9}$	$19.89_{3.3}$	$14.12_{0.9}$	$26.83_{2.8}$	$19.32_{1.2}$	$\mathbf{43.64}_{2.5}$

CoLA and DBpedia. FedRespro ranks first on both encoder tasks (Table 2), reaching 43.00 CoLA MCC ($\times 100$) and 43.64% DBpedia accuracy. These results improve on RB-LoRA, the strongest baseline on both tasks, by 23.43 MCC points and 16.81 percentage points, respectively. On CoLA, FedRespro also reduces the standard deviation from RB-LoRA's 11.3 to 0.8. The combination of larger mean scores and more consistent CoLA results demonstrates the benefits of residual collaboration across both grammatical acceptability and topic classification.

4.2 Ablation Studies

Qwen2 variants. Full FedRespro achieves the highest server mean at all three scales in Table 3(a). Removing downlink memory lowers accuracy by 0.55, 1.33, and 1.47 percentage points at 0.5B, 1.5B, and 7B, respectively. Global correction contributes most at 1.5B, where its removal reduces accuracy from 63.23% to 59.07%; the corresponding drops at 0.5B and 7B are 0.68 and 0.28 points. Both components therefore improve server performance, with global correction particularly beneficial at 1.5B and downlink memory contributing more at 7B.

BERT-base variants. Cumulative residual history provides the largest gains on both encoder tasks (Table 3(b)). Using only the current round's residual reduces CoLA from 43.00 to 15.35 MCC and DBpedia from 43.64% to 14.07%, losses of 27.65 MCC points and 29.57 percentage points.

Downlink memory also supports more consistent outcomes: its removal fails on CoLA and increases DBpedia's standard deviation from 2.51 to 11.56 despite a similar mean accuracy. These results highlight the value of preserving residual knowledge across rounds for both predictive performance and stability.

Table 3: FedRespro server ablations. (a) Qwen2 eight-task accuracy (%); Δ is full minus w/o downlink memory at 0.5B/1.5B and w/o global correction at 7B. (b) BERT-base CoLA MCC $\times 100$ and DBpedia accuracy (%); Δ is full minus round-only. Subscripts give three-seed sample SD; paired differences are computed before rounding, while DBpedia Δ uses the reported means. "failed" denotes failure to converge.

(a) Qwen2

Configuration	0.5B	1.5B	7B
FedRespro	$41.24_{3.63}$	$63.23_{1.18}$	$79.84_{0.07}$
w/o downlink memory	$40.69_{3.11}$	$61.91_{0.90}$	$78.37_{0.07}$
w/o global correction	$40.56_{1.03}$	$59.07_{0.73}$	$79.57_{0.17}$
Δ	$+0.55_{4.51}$	$+1.33_{0.78}$	$+0.28_{0.23}$

(b) BERT-base

Configuration	CoLA	DBpedia
FedRespro	$43.00_{0.83}$	$43.64_{2.51}$
w/o downlink memory	failed	$43.57_{11.56}$
Round-only residual	$15.35_{4.13}$	$14.07_{4.46}$
Δ (round-only)	$+27.65_{3.53}$	$+29.57$

4.3 Hyperparameter Sensitivity and Communication Efficiency

Table 4: Sensitivity and communication over 40 rounds. (a) Qwen2-0.5B server accuracy in a matched λ sweep, seed 0, without checkpoint averaging. (b) Qwen2-1.5B/7B upload plus download in GiB, including initialization. Subscripts give three-seed sample SD.

(a) Sensitivity to λ

λ	Acc. (%)	λ	Acc. (%)
0.5	39.15	1.1	41.95
0.9	43.00	2	39.08
1	44.30	4	37.90

(b) Communication (GiB)

Method	1.5B ↓	7B ↓	Method	1.5B ↓	7B ↓
HetLoRA	$0.876_{0.013}$	$1.972_{0.040}$	RB-LoRA	$0.921_{0.009}$	$2.134_{0.021}$
FlexLoRA	$0.921_{0.009}$	$2.134_{0.021}$	FSLoRA	$0.797_{0.003}$	$1.845_{0.007}$
FLoRA	$4.240_{0.024}$	$9.819_{0.057}$	**FedRespro**	$0.921_{0.009}$	$2.134_{0.021}$

Sensitivity to λ. We vary the weight of complementary residuals while fixing the seed, client schedule, and transmission rank (Table 4(a)). At 0.5B, $\lambda = 1$ achieves the highest observed accuracy (44.30%); 0.9/1.1 give 43.00/41.95%, while 0.5, 2, and 4 give 39.15%, 39.08%, and 37.90%, respectively. At 1.5B and 7B, $\lambda = 2$ exceeds 2.25 and 1.5 by 1.63 and 1.88 points, respectively. These single-seed checks suggest scale-dependent sensitivity; matched protocols are provided in the appendix.

Communication efficiency. Each client receives the global LoRA projected to its own rank (4 or 8), while residual memory remains at the server. Over 40 rounds, FedRespro uses mean totals of 0.921 and 2.134 GiB at 1.5B and 7B, respectively (Table 4(b)). These volumes are 78.3% lower than FLoRA at both scales. This adapter-only communication configuration differs from the main performance configurations and from the separate Residual-Sketch channel analyzed in Appendix C.2. The communication totals therefore do not establish a joint accuracy–communication operating point for Table 1. Appendix H provides the accounting details.

5 Conclusion

In this work, we identified the Resource Inverse Distribution (RID) dilemma in heterogeneous model collaboration and proposed **FedRespro** to recouple statistical reliability and optimization capacity across clients. FedRespro discovers reliable complementary knowledge through spatial–temporal residual assessment and preserves it in a dense cross-round state, while capacity-aware rank-matched projection enables efficient knowledge transfer to resource-constrained clients. We further establish theoretical guarantees for reliability estimation and convergence. Extensive experiments demonstrate the effectiveness of FedRespro across diverse RID settings, with substantial performance gains while avoiding dense residual communication.

AI USE STATEMENT

AI-assisted tools were used for language editing, manuscript integration, formatting, and consistency checks on author-provided experimental summaries. The authors are responsible for the scientific content and for validating the underlying experiments, analysis, and conclusions.

REFERENCES

Samiul Alam, Luyang Liu, Ming Yan, and Mi Zhang. FedRolex: Model-heterogeneous federated learning with rolling sub-model extraction. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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*, volume 37, 2024.

Hanchi Chen and Haris Vikalo. Heterogeneity-guided client sampling: Towards fast and efficient non-IID federated learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65525–65561, 2024.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish Khisti. Robust federated finetuning of LLMs via alternating optimization of LoRA. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yujun Cheng, Weiting Zhang, Zhewei Zhang, Chuan Zhang, Shengjin Wang, and Shiwen Mao. Toward federated large language models: Motivations, methods, and future directions. *IEEE Communications Surveys & Tutorials*, 27(4):2733–2764, 2025. doi: 10.1109/COMST.2024.3503680.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous LoRA for federated fine-tuning of on-device foundation models. *arXiv preprint arXiv:2401.06432*, 2024. URL https://arxiv.org/abs/2401.06432.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2089–2099, 2021.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Enmao Diao, Jie Ding, and Vahid Tarokh. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2021.

Haizhou Du, Yiran Xiang, Yiwen Cai, Xiufeng Liu, Zonghan Wu, Huan Huo, and Guodong Long. FedFree: Breaking knowledge-sharing barriers through layer-wise alignment in heterogeneous federated learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

John C. Duchi, Vitaly Feldman, Lunjia Hu, and Kunal Talwar. Subspace recovery from heterogeneous data with non-isotropic noise. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5854–5866, 2022.

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan, Seyyedali Hosseinalipour, and Christopher G. Brinton. Federated sketching LoRA: A flexible framework for heterogeneous collaborative fine-tuning of LLMs. *arXiv preprint arXiv:2501.19389*, 2025. URL https://arxiv.org/abs/2501.19389. Version 4, revised May 2026.

Sihyeon Ha, Yongjeong Oh, and Yo-Seb Jeon. RB-LoRA: Rank-balanced aggregation for low-rank adaptation with federated fine-tuning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1737–1746, 2026. doi: 10.18653/v1/2026.findings-eacl.88. URL https://aclanthology.org/2026.findings-eacl.88/.

Fatih Ilhan, Gong Su, and Ling Liu. ScaleFL: Resource-adaptive federated learning with heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24532–24541, 2023.

Wonyong Jeong and Sung Ju Hwang. Factorized-FL: Personalized federated learning with parameter factorization & similarity matching. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yongzhe Jia, Xuyun Zhang, Hongsheng Hu, Kim-Kwang Raymond Choo, Lianyong Qi, Xiaolong Xu, Amin Beheshti, and Wanchun Dou. DapperFL: Domain adaptive federated learning with model fusion pruning for edge devices. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5132–5143, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-iid data. In *International Conference on Learning Representations*, 2020.

Jiahao Liu, Yipeng Zhou, Di Wu, Miao Hu, Mohsen Guizani, and Quan Z. Sheng. FedLMT: Tackling system heterogeneity of federated learning via low-rank model training with theoretical guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32509–32551, 2024.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 21884–21914, 2023.

Kangyang Luo, Shuai Wang, Yexuan Fu, Xiang Li, Yunshi Lan, and Ming Gao. DFRD: Data-free robustness distillation for heterogeneous federated learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Chuiyang Meng, Ming Tang, and Vincent Wong. FLoRG: Federated fine-tuning with low-rank gram matrices and procrustes alignment. In *International Conference on Learning Representations*, 2026.

Arian Raje, Baris Askin, Divyansh Jhunjhunwala, and Gauri Joshi. Ravan: Multi-head low-rank adaptation for federated fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations*, 2024.

Jiaqi Wang, Xingyi Yang, Suhan Cui, Liwei Che, Lingjuan Lyu, Dongkuan Xu, and Fenglong Ma. Towards personalized federated learning via heterogeneous model reassembly. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jiaqi Wang, Chen Zhao, Lingjuan Lyu, Quanzeng You, Mengdi Huai, and Fenglong Ma. Bridging model heterogeneity in federated learning via uncertainty-based asymmetrical reciprocity learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Lei Wang, Jieming Bian, Letian Zhang, and Jie Xu. Adaptive LoRA experts allocation and selection for federated fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972. doi: 10.1007/BF01932678.

K.-S. Wong, M. Nguyen-Duc, K. Le, L. Ho, et al. An empirical study of federated learning on IoT–edge devices: Resource allocation and heterogeneity. *IEEE Transactions on Neural Networks and Learning Systems*, 37(2):753–765, 2026.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Efficient federated learning against heterogeneous and non-stationary client unavailability. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Rick Siow Mong Goh, Yong Liu, and Lei Zhu. Federated residual low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2025.

Kai Yi, Nidham Gazagnadou, Peter Richtárik, and Lingjuan Lyu. FedP3: Federated personalized and privacy-friendly network pruning under model heterogeneity. In *International Conference on Learning Representations*, 2024a.

Liping Yi, Han Yu, Chao Ren, Gang Wang, Xiaoxiao Liu, and Xiaoxiao Li. Federated model heterogeneous matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-2125.

A THEORETICAL DETAILS OF FEDRESPRO

This section details the FedRespro procedure and explains why the spatial consistency measured in the statistically supported low-dimensional subspace can provide reliability evidence for the complementary residual of a high-capacity client. We first specify the required assumptions, then describe the reference-aligned projection, and finally establish the approximation between shared-space disagreement and residual uncertainty.

A.1 COMPLETE FEDRESPRO PROCEDURE

Algorithm 1 FedRespro with Residual-Sketch Transport

Require: Client ranks $\{r_k\}$, local data sizes $\{n_k\}$, complementary coefficient $\{\lambda_t\}$

Initialize shared state $\mathbf{W}^0$ and dense complementary state $\mathbf{M}^0 = \mathbf{0}$.
for $t = 1, \dots, T$ **do**
  Sample participating clients $\mathcal{K}_t$.
  for each $k \in \mathcal{K}_t$ **do**
    Construct $\mathbf{S}_k^{t-1} = \mathcal{P}_{r_k}(\mathbf{M}^{t-1})$ and transmit its factorized Residual-Sketch.
    Incorporate $\mathbf{S}_k^{t-1}$ into the round-specific optimization base while keeping $\Delta\mathbf{W}_k^{\mathrm{ref}}$ fixed.
    Client performs standard rank-r_k local optimization initialized from $\Delta\mathbf{W}_k^{\mathrm{ref}}$ and returns $\Delta\mathbf{W}_{k,t}$.
    Server reconstructs $\mathbf{G}_k^t = \Delta\mathbf{W}_{k,t} - \Delta\mathbf{W}_k^{\mathrm{ref}}$.
  end for
  Construct the statistical reference $\mathbf{G}_{\mathcal{L}}^t$ and complementary residuals $\{\mathbf{R}_k^t\}_{k\in\mathcal{H}}$.
  Compute spatial consistency $\{s_k^t\}$.
  if $t = 1$ or $\|\mathbf{M}^{t-1}\|_F \le \epsilon_M$ **then**
    Set $p_k^t = 1$ for all $k \in \mathcal{H}$.
  else
    Compute p_k^t according to Eq. 8.
  end if
  Compute $Z_t = \sum_{k\in\mathcal{H}} \sqrt{s_k^t p_k^t}$.
  if $Z_t > 0$ **then**
    $\beta_k^t = \sqrt{s_k^t p_k^t}/Z_t$.
    $\mathbf{R}_{\mathrm{rel}}^t = \sum_{k\in\mathcal{H}} \beta_k^t \mathbf{R}_k^t$.
  else
    Set $\mathbf{R}_{\mathrm{rel}}^t = \mathbf{0}$.
  end if
  Form $\mathbf{G}^t = \mathbf{G}_{\mathcal{L}}^t + \lambda_t \mathbf{R}_{\mathrm{rel}}^t$.
  Update

$$\mathbf{W}^t = \mathbf{W}^{t-1} + \mathbf{G}_{\mathcal{L}}^t, \qquad \mathbf{M}^t = \mathbf{M}^{t-1} + \lambda_t \mathbf{R}_{\mathrm{rel}}^t.$$

end for

The complete procedure of the proposed FedRespro is summarized in **Algorithm 1**. For analysis, we define the effective dense server state as $\mathbf{X}^t = \mathbf{W}^t + \mathbf{M}^t$. The two-state update above immediately gives $\mathbf{X}^t - \mathbf{X}^{t-1} = \mathbf{G}^t$, which recovers Eq. 12. The convergence analysis below reindexes rounds from zero. Meanwhile, each client accesses only the rank-matched approximation $\mathbf{W}^t + \mathcal{P}_{r_k}(\mathbf{M}^t)$. The residual memory $\mathbf{M}^t$ does not alter the fixed low-rank reference $\Delta\mathbf{W}_k^{\mathrm{ref}}$. Instead, its rank-matched Residual-Sketch is incorporated only into the round-specific optimization base. Hence, residuals from different rounds retain a consistent reference semantics. Finally, we use stratified participation so that $\mathcal{L}_t \neq \emptyset$ whenever the reliability branch is evaluated; if $\mathcal{H}_t = \emptyset$, the complementary update is skipped.

A.2 ASSUMPTIONS AND ANALYSIS SETTING

Our analysis builds on standard modeling conditions commonly adopted in model-heterogeneous federated learning, stochastic optimization, and subspace estimation. In particular, heterogeneous low-rank models are typically embedded into a common ambient parameter space (Bai et al., 2024;

Liu et al., 2024), while shared and client-specific components have been widely used to characterize heterogeneous federated representations (Collins et al., 2021; Jeong & Hwang, 2022). The stochastic-noise and spectral-separation conditions below follow standard assumptions in federated optimization and matrix perturbation analysis (Li et al., 2020; Karimireddy et al., 2020; Wedin, 1972). We further allow client-dependent anisotropic noise, consistent with recent subspace-recovery analyses under heterogeneous federated data (Duchi et al., 2022).

Assumption 1 (Aligned and nested update spaces) *Following heterogeneous low-rank FL formulations (Bai et al., 2024; Liu et al., 2024), all client updates are represented in a common ambient parameter space $\mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Let $\mathbf{U}_{\mathcal{L}}, \mathbf{V}_{\mathcal{L}}$ denote orthonormal bases of the rank-$r_{\mathcal{L}}$ reference-supported left and right subspaces, respectively. For each high-capacity client k with $r_k > r_{\mathcal{L}}$, let $\mathbf{U}_k, \mathbf{V}_k$ denote the corresponding rank-r_k accessible bases. We assume*

$$\operatorname{span}(\mathbf{U}_{\mathcal{L}}) \subseteq \operatorname{span}(\mathbf{U}_k), \qquad \operatorname{span}(\mathbf{V}_{\mathcal{L}}) \subseteq \operatorname{span}(\mathbf{V}_k). \tag{18}$$

Assumption 2 (Shared complementary decomposition) *Following the shared–client-specific decomposition commonly adopted in heterogeneous and personalized federated learning (Collins et al., 2021; Jeong & Hwang, 2022), we assume that the signal component of a high-capacity update can be decomposed into a statistically supported shared component and an additional client-dependent component:*

$$\mathbf{G}_k = \mathbf{G}_\star + \mathbf{C}_k + \mathbf{E}_k, \tag{19}$$

where $\mathbf{G}_\star$ denotes the shared optimization component and $\mathbf{E}_k$ denotes finite-sample estimation noise.

Let $\mathcal{P}_\star$ denote the ideal projection onto the subspace spanned by $\mathbf{G}_\star$. We define the complementary component as the portion of the signal outside this shared subspace. Consequently,

$$\mathcal{P}_\star(\mathbf{C}_k) = \mathbf{0}, \tag{20}$$

holds by construction rather than as an additional orthogonality assumption.

Assumption 3 (Finite sample estimation noise) *The local update is treated as a stochastic estimate of its underlying optimization signal. Following standard stochastic and federated optimization analyses (Li et al., 2020; Karimireddy et al., 2020), we assume that the estimation error is zero-mean and has bounded second moment:*

$$\mathbb{E}[\mathbf{E}_k] = \mathbf{0}, \qquad \mathbb{E}\|\mathbf{E}_k\|_F^2 \le \frac{\sigma_k^2}{n_k}, \tag{21}$$

where n_k denotes the effective number of independent local observations. For the data-rich clients used to construct the statistical reference, we additionally assume that their estimation errors are conditionally independent across clients. The $1/n_k$ dependence reflects the standard variance reduction obtained by averaging independent stochastic observations.

Assumption 4 (Stable reference subspace) *Let the statistical reference constructed from data-rich clients satisfy*

$$\mathbf{G}_{\mathcal{L}} = \mathbf{G}_\star + \mathbf{E}_{\mathcal{L}}, \tag{22}$$

and denote

$$\varepsilon_{ref} = \|\mathbf{E}_{\mathcal{L}}\|_2. \tag{23}$$

We assume that the statistically supported component exhibits a non-vanishing singular-value gap at rank r:

$$\gamma = \sigma_{r_{\mathcal{L}}}(\mathbf{G}_\star) - \sigma_{r_{\mathcal{L}}+1}(\mathbf{G}_\star) > 0, \tag{24}$$

with

$$\varepsilon_{ref} < \gamma. \tag{25}$$

Such a spectral-separation condition is standard in singular-subspace perturbation analysis and ensures that the leading reference subspace remains identifiable under finite-sample perturbations (Wedin, 1972; Davis & Kahan, 1970).

Assumption 5 (Bounded anisotropic client-level noise) *Let*

$$\mathbf{\Sigma}_k = \mathrm{Cov}\left[\mathrm{vec}(\mathbf{E}_k)\right]. \tag{26}$$

Following heterogeneous subspace-recovery analyses with non-isotropic noise (Duchi et al., 2022), we allow client-dependent covariance and assume that there exist constants $0 < c_- \le c_+ < \infty$ such that

$$c_- \frac{\tau_k^2}{n_k}\mathbf{I} \preceq \mathbf{\Sigma}_k \preceq c_+ \frac{\tau_k^2}{n_k}\mathbf{I}. \tag{27}$$

We define $\kappa_\Sigma = c_+/c_-$. Thus, the noise may be client-dependent and anisotropic, while its energy over any fixed subspace remains controlled by the common finite-sample scale τ_k^2/n_k.

For the high-probability analysis, following standard stochastic optimization analyses (Liu et al., 2023), we additionally assume uniformly sub-Gaussian projected noise: for every unit vector $\mathbf{u}$,

$$\left\|\langle \mathbf{u}, \mathrm{vec}(\mathbf{E}_k)\rangle\right\|_{\psi_2} \le \kappa_{\mathrm{sg}} \frac{\tau_k}{\sqrt{n_k}}, \tag{28}$$

where $\kappa_{\mathrm{sg}} < \infty$ is a uniform sub-Gaussian parameter. **Reference-Aligned Low-Dimensional Projection**

FedRespro first constructs a statistically supported reference from the data-rich clients:

$$\mathbf{G}_{\mathcal{L}} = \sum_{j\in\mathcal{L}} \frac{n_j}{N_{\mathcal{L}}}\mathbf{G}_j, \qquad N_{\mathcal{L}} = \sum_{j\in\mathcal{L}} n_j. \tag{29}$$

Under **Assumption 3**, its estimation uncertainty decreases with $N_{\mathcal{L}}$, making it a statistically grounded optimization reference.

Let

$$\mathbf{G}_{\mathcal{L}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top, \qquad \bar{\mathbf{G}}_{\mathcal{L}} = \widehat{\mathbf{U}}_r\widehat{\mathbf{\Sigma}}_r\widehat{\mathbf{V}}_r^\top, \tag{30}$$

where the leading r singular directions define the statistically supported low-dimensional subspace. Rather than independently truncating each high-capacity update, we project it into this reference-aligned coordinate system:

$$\widetilde{\mathbf{G}}_k = \widehat{\mathbf{U}}_r^\top \mathbf{G}_k \widehat{\mathbf{V}}_r, \qquad \mathbf{R}_k = \mathbf{G}_k - \widehat{\mathbf{U}}_r\widetilde{\mathbf{G}}_k\widehat{\mathbf{V}}_r^\top. \tag{31}$$

Since $\widehat{\mathbf{U}}_r$ and $\widehat{\mathbf{V}}_r$ are orthonormal, the shared-space alignment can be equivalently evaluated in the $r \times r$ coordinate space, where the reference is $\widehat{\mathbf{\Sigma}}_r$:

$$s_k = \left[\frac{\left\langle \widetilde{\mathbf{G}}_k, \widehat{\mathbf{\Sigma}}_r\right\rangle_F}{\|\widetilde{\mathbf{G}}_k\|_F\|\widehat{\mathbf{\Sigma}}_r\|_F + \epsilon}\right]_+. \tag{32}$$

Thus, s_k measures the consistency of the high-capacity update with the statistically supported optimization tendency, while $\mathbf{R}_k$ retains its complementary component outside the reference subspace.

A.3 From Low-Dimensional Similarity to Residual Reliability

We next explain how the similarity in Eq. 32 provides information about the statistical uncertainty of $\mathbf{R}_k$. Let

$$a_k = \|\widetilde{\mathbf{G}}_k\|_F, \qquad b = \|\widehat{\mathbf{\Sigma}}_r\|_F. \tag{33}$$

Their squared discrepancy is exactly

$$\begin{aligned} d_{k,\|}^2 &= \|\widetilde{\mathbf{G}}_k - \widehat{\mathbf{\Sigma}}_r\|_F^2 \\ &= (a_k - b)^2 + 2a_k b(1 - \bar{s}_k). \end{aligned} \tag{34}$$

Hence, the spatial similarity is directly related to the disagreement observed in the statistically supported subspace,

$$d_{k,\|}^2 \approx \|\widehat{\mathcal{P}}(\mathbf{E}_k)\|_F^2, \tag{35}$$

where $\widehat{\mathcal{P}}$ denotes the reference-aligned projection. Under **Assumption 1**, let

$$q_{\mathcal{L}} = r_{\mathcal{L}}^2, \qquad q_{k,\perp} = r_k^2 - r_{\mathcal{L}}^2 \tag{36}$$

denote the effective dimensions of the shared and accessible complementary subspaces, respectively. Let $\mathcal{P}_{k,\perp}$ denote the orthogonal projection onto $\mathcal{S}_{k,\perp}$. For any orthogonal projector onto a q-dimensional subspace, **Assumption 5** implies that its projected noise energy is bounded by the trace of the corresponding covariance restriction. Hence,

$$c_- q_{\mathcal{L}} \frac{\tau_k^2}{n_k} \le \mathbb{E}\|\mathcal{P}_{\mathcal{L}}(\mathbf{E}_k)\|_F^2 \le c_+ q_{\mathcal{L}} \frac{\tau_k^2}{n_k}, \tag{37}$$

and

$$c_- q_{k,\perp} \frac{\tau_k^2}{n_k} \le \mathbb{E}\|\mathcal{P}_{k,\perp}(\mathbf{E}_k)\|_F^2 \le c_+ q_{k,\perp} \frac{\tau_k^2}{n_k}. \tag{38}$$

Therefore,

$$\frac{1}{\kappa_\Sigma} \frac{q_{k,\perp}}{q_{\mathcal{L}}} \le \frac{\mathbb{E}\|\mathcal{P}_{k,\perp}(\mathbf{E}_k)\|_F^2}{\mathbb{E}\|\mathcal{P}_{\mathcal{L}}(\mathbf{E}_k)\|_F^2} \le \kappa_\Sigma \frac{q_{k,\perp}}{q_{\mathcal{L}}}. \tag{39}$$

In the isotropic special case $c_- = c_+$, Eq. 39 reduces to

$$\frac{\mathbb{E}\|\mathcal{P}_{k,\perp}(\mathbf{E}_k)\|_F^2}{\mathbb{E}\|\mathcal{P}_{\mathcal{L}}(\mathbf{E}_k)\|_F^2} = \frac{r_k^2 - r_{\mathcal{L}}^2}{r_{\mathcal{L}}^2}. \tag{40}$$

Together with the normalized shared-space disagreement $2(1 - s_k^t)$, this motivates

$$\widehat{\mathcal{U}}_{k,\perp}^t = \frac{r_k^2 - r_{\mathcal{L}}^2}{r_{\mathcal{L}}^2} 2(1 - s_k^t). \tag{41}$$

Thus, the estimator is exact in dimensional scaling under isotropic noise and remains a constant-factor reliability proxy under bounded anisotropy.

A.4 Approximation Analysis of Complementary Residual Reliability

We provide the derivation of **Theorem 1** and **Corollary 1** in the section. Let

$$q_{\mathcal{L}} = r_{\mathcal{L}}^2, \qquad q_{k,\perp} = r_k^2 - r_{\mathcal{L}}^2 \tag{42}$$

denote the effective dimensions of the shared and complementary subspaces, respectively. Recall that the reference-aligned representation of client k and the statistical reference are $\widetilde{\mathbf{G}}_k^t$ and $\widehat{\mathbf{\Sigma}}_{r_{\mathcal{L}}}^t$. Define their normalized forms as

$$\mathbf{A}_k^t = \frac{\widetilde{\mathbf{G}}_k^t}{\|\widetilde{\mathbf{G}}_k^t\|_F}, \qquad \mathbf{A}_{\mathcal{L}}^t = \frac{\widehat{\mathbf{\Sigma}}_{r_{\mathcal{L}}}^t}{\|\widehat{\mathbf{\Sigma}}_{r_{\mathcal{L}}}^t\|_F}. \tag{43}$$

For $s_k^t > 0$, the clipping operation is inactive, and therefore

$$\begin{aligned} d_{k,\|}^{t\,2} &= \|\mathbf{A}_k^t - \mathbf{A}_{\mathcal{L}}^t\|_F^2 \\ &= 2 - 2\left\langle \mathbf{A}_k^t, \mathbf{A}_{\mathcal{L}}^t \right\rangle_F = 2(1 - s_k^t). \end{aligned} \tag{44}$$

Hence, $2(1 - s_k^t)$ is exactly the normalized directional disagreement observed in the statistically supported subspace.

Under **Assumption 5**, the estimation noise in the shared and complementary subspaces is governed by the same client-level uncertainty scale, while their total noise energies differ primarily with the corresponding effective dimensions. In the isotropic case,

$$\frac{\mathbb{E}\|\mathbf{E}_{k,\perp}^t\|_F^2}{\mathbb{E}\|\mathbf{E}_{k,\|}^t\|_F^2} = \frac{q_{k,\perp}}{q_{\mathcal{L}}} = \frac{r_k^2 - r_{\mathcal{L}}^2}{r_{\mathcal{L}}^2}. \tag{45}$$

The bounded-anisotropy condition in **Assumption 5** relaxes this equality to a constant-factor approximation without changing its dimensional scaling. Combining Eqs. 44 and 45 motivates the estimator used in **Theorem 1**:

$$\widehat{\mathcal{U}}_{k,\perp}^t = \frac{r_k^2 - r_{\mathcal{L}}^2}{r_{\mathcal{L}}^2} 2(1 - s_k^t). \tag{46}$$

Thus, FedRespro uses the statistically observable disagreement in the shared subspace to infer the uncertainty of directions that are only accessible to the high-capacity client.

The deviation between $\widehat{\mathcal{U}}_{k,\perp}^t$ and the latent complementary uncertainty $\mathcal{U}_{k,\perp}^t$ contains three components: statistical-reference estimation, shared-subspace concentration, and complementary-subspace concentration. Under **Assumptions 3–5**, there exist finite constants C_{ref}, $C_{\|}$, and $C_{\perp}$ such that, for any $\delta \in (0,1)$,

$$\Pr\left(\mathcal{E}_{\mathrm{ref}} > C_{\mathrm{ref}}\sqrt{\frac{\log(6/\delta)}{N_{\mathcal{L}}}}\right) \le \frac{\delta}{3}, \tag{47}$$

$$\Pr\left(\mathcal{E}_{\|} > C_{\|}\frac{\sqrt{\log(6/\delta)}}{r_{\mathcal{L}}}\right) \le \frac{\delta}{3}, \tag{48}$$

$$\Pr\left(\mathcal{E}_{\perp} > C_{\perp}\frac{\sqrt{\log(6/\delta)}}{\sqrt{r_k^2 - r_{\mathcal{L}}^2}}\right) \le \frac{\delta}{3}. \tag{49}$$

The constants depend only on the bounded-anisotropy, sub-Gaussian-noise, and stable-subspace parameters. Define

$$C_{\mathrm{rel}} = \max\{C_{\mathrm{ref}}, C_{\|}, C_{\perp}\}. \tag{50}$$

Applying a union bound yields, with probability at least $1-\delta$,

$$\left\|\widehat{\mathcal{U}}_{k,\perp}^t - \mathcal{U}_{k,\perp}^t\right\| \le C_{\mathrm{rel}}\sqrt{\log\frac{6}{\delta}}\left(N_{\mathcal{L}}^{-1/2} + r_{\mathcal{L}}^{-1} + (r_k^2 - r_{\mathcal{L}}^2)^{-1/2}\right), \tag{51}$$

which proves **Corollary 1**.

B Convergence Analysis of FedRespro

We establish the convergence of FedRespro by viewing its server update as an inexact stochastic descent step. The analysis explicitly characterizes the errors introduced by reliability-aware complementary knowledge discovery and capacity-aware rank-matched residual projection.

B.1 Optimization Objective and Assumptions

Let the global objective be

$$F(\mathbf{W}) = \sum_{k=1}^{K} p_k F_k(\mathbf{W}), \qquad \sum_{k=1}^{K} p_k = 1, \tag{52}$$

where F_k denotes the local objective of client k and p_k is its aggregation weight.

FedRespro maintains two non-overlapping server states: the shared state $\mathbf{W}^t$ and the complementary residual state $\mathbf{M}^t$. For analysis, define the effective dense state

$$\mathbf{X}^t = \mathbf{W}^t + \mathbf{M}^t. \tag{53}$$

Since

$$\mathbf{W}^{t+1} = \mathbf{W}^t + \mathbf{G}_{\mathcal{L}}^t, \qquad \mathbf{M}^{t+1} = \mathbf{M}^t + \lambda_t \mathbf{R}_{\mathrm{rel}}^t,$$

we have

$$\mathbf{X}^{t+1} - \mathbf{X}^t = \mathbf{G}_{\mathcal{L}}^t + \lambda_t \mathbf{R}_{\mathrm{rel}}^t = \mathbf{G}^t.$$

For convergence analysis, we express the effective FedRespro update as an inexact stochastic descent step on $\mathbf{X}^t$:

$$\mathbf{X}^{t+1} = \mathbf{X}^t - \eta_t \widehat{\mathbf{g}}^t, \qquad \widehat{\mathbf{g}}^t = -\frac{\mathbf{G}^t}{\eta_t}. \tag{54}$$

Here η_t denotes the effective round-level step size induced by the local optimizer and server update, so that $\mathbf{G}^t = -\eta_t \widehat{\mathbf{g}}^t$. We decompose

$$\widehat{\mathbf{g}}^t = \nabla F(\mathbf{X}^t) + \boldsymbol{\xi}^t + \mathbf{b}_{\mathrm{mem}}^t + \mathbf{b}_{\mathrm{rank}}^t + \mathbf{b}_{\mathrm{loc}}^t, \tag{55}$$

where $\boldsymbol{\xi}^t$ denotes stochastic gradient noise, while $\mathbf{b}^t_{\mathrm{mem}}$, $\mathbf{b}^t_{\mathrm{rank}}$, and $\mathbf{b}^t_{\mathrm{loc}}$ capture the errors induced by imperfect reliability-aware memory construction, capacity-constrained projection, and local optimization, respectively.

Assumption 6 (Smoothness and lower boundedness) *Each F_k is L-smooth, i.e.,*

$$\|\nabla F_k(\mathbf{W}) - \nabla F_k(\mathbf{W}')\|_F \le L\|\mathbf{W} - \mathbf{W}'\|_F, \tag{56}$$

and $F(\mathbf{W})$ is lower bounded by $F^\star$.

Assumption 7 (Stochastic gradient noise) *The stochastic component satisfies*

$$\mathbb{E}[\boldsymbol{\xi}^t \mid \mathbf{W}^t] = \mathbf{0}, \qquad \mathbb{E}\|\boldsymbol{\xi}^t\|_F^2 \le \sigma^2. \tag{57}$$

Assumption 8 (Bounded client heterogeneity) *The local gradients satisfy*

$$\sum_{k=1}^{K} p_k \|\nabla F_k(\mathbf{W}) - \nabla F(\mathbf{W})\|_F^2 \le \zeta^2. \tag{58}$$

We additionally use the regularity conditions established in Appendix A for spatial consistency and assume that the normalization term in the reliability-aware aggregation remains bounded away from zero. This is a uniform positive lower-bound condition on active-round normalizers, which is stronger than requiring non-zero evidence in each round.

B.2 Error Induced by Reliable Complementary Knowledge Discovery

We first characterize how imperfect spatial and temporal reliability estimation affects the accumulated complementary residual state.

Let $s_k^{t,\star}$ and $p_k^{t,\star}$ denote the ideal spatial and temporal reliability scores of client k, respectively, and define

$$a_k^t = s_k^t p_k^t, \qquad a_k^{t,\star} = s_k^{t,\star} p_k^{t,\star}. \tag{59}$$

The corresponding ideal normalized reliability weight is

$$\beta_k^{t,\star} = \frac{\sqrt{a_k^{t,\star}}}{Z_t^\star}, \qquad Z_t^\star = \sum_{h\in\mathcal{H}} \sqrt{a_h^{t,\star}}, \tag{60}$$

and the ideal complementary residual is

$$\mathbf{R}_{\mathrm{rel}}^{t,\star} = \sum_{k\in\mathcal{H}} \beta_k^{t,\star} \mathbf{R}_k^t. \tag{61}$$

Spatial and temporal score errors. Let

$$c_k = \frac{r_k^2 - r_{\mathcal{L}}^2}{r_{\mathcal{L}}^2}. \tag{62}$$

Since $\widehat{\mathcal{U}}_{k,\perp}^t = 2c_k(1 - s_k^t)$, **Corollary 1** implies, on its $1-\delta$ event,

$$|s_k^t - s_k^{t,\star}| \le \Delta_{k,s}^t(\delta), \qquad \Delta_{k,s}^t(\delta) = \frac{r_{\mathcal{L}}^2}{2(r_k^2 - r_{\mathcal{L}}^2)} \Delta_k^t(\delta). \tag{63}$$

For the temporal score, we denote

$$|p_k^t - p_k^{t,\star}| \le \Delta_{k,p}^t \tag{64}$$

and assume that $\Delta_{k,p}^t$ is bounded. This condition follows from the stability of normalized cosine similarity whenever the compared residual norms are bounded away from zero.

Because $s_k^t, p_k^t \in [0,1]$,

$$|a_k^t - a_k^{t,\star}| \le \Delta_{k,s}^t + \Delta_{k,p}^t. \tag{65}$$

Residual aggregation error. Assume that the active-round normalizers satisfy

$$Z_t \geq z_{\min} > 0, \qquad Z_t^\star \geq z_{\min} > 0, \tag{66}$$

and that

$$\|\mathbf{R}_k^t\|_F \leq G_R \qquad \forall k, t. \tag{67}$$

Using $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$ for $x, y \geq 0$, the normalized geometric weights satisfy

$$\sum_{k\in\mathcal{H}} |\beta_k^t - \beta_k^{t,\star}| \leq \frac{2}{z_{\min}} \sum_{k\in\mathcal{H}} \sqrt{\Delta_{k,s}^t + \Delta_{k,p}^t}. \tag{68}$$

Consequently,

$$\left\|\mathbf{R}_{\mathrm{rel}}^t - \mathbf{R}_{\mathrm{rel}}^{t,\star}\right\|_F^2 \leq \frac{4G_R^2}{z_{\min}^2} \left(\sum_{k\in\mathcal{H}} \sqrt{\Delta_{k,s}^t + \Delta_{k,p}^t} \right)^2. \tag{69}$$

We denote the right-hand side by $\varepsilon_{\mathrm{rel},t}^2$.

When $Z_t = 0$, FedRespro skips the complementary branch and sets $\mathbf{R}_{\mathrm{rel}}^t = \mathbf{0}$, as specified in Eq. 10; hence no numerically unstable normalization is performed.

Accumulated memory error. Let the ideal dense complementary state evolve as

$$\mathbf{M}_\star^{t+1} = \mathbf{M}_\star^t + \lambda_t \mathbf{R}_{\mathrm{rel}}^{t,\star}, \qquad \mathbf{M}_\star^0 = \mathbf{0}. \tag{70}$$

The actual FedRespro state satisfies

$$\mathbf{M}^{t+1} = \mathbf{M}^t + \lambda_t \mathbf{R}_{\mathrm{rel}}^t. \tag{71}$$

Therefore,

$$\begin{aligned} \mathbf{M}^{t+1} - \mathbf{M}_\star^{t+1} = \mathbf{M}^t - \mathbf{M}_\star^t \\ + \lambda_t \left(\mathbf{R}_{\mathrm{rel}}^t - \mathbf{R}_{\mathrm{rel}}^{t,\star} \right). \end{aligned} \tag{72}$$

Define

$$\varepsilon_{\mathrm{mem},t} = \left(\mathbb{E}\|\mathbf{M}^t - \mathbf{M}_\star^t\|_F^2 \right)^{1/2}. \tag{73}$$

By Minkowski's inequality,

$$\varepsilon_{\mathrm{mem},t} \leq \sum_{\tau=0}^{t-1} |\lambda_\tau| \varepsilon_{\mathrm{rel},\tau}. \tag{74}$$

Thus, reliability-estimation error affects optimization through the cross-round complementary state rather than as an independent instantaneous bias.

B.3 Error Induced by Rank-Matched Residual Projection

FedRespro preserves the complementary knowledge in the dense server-side state $\mathbf{M}^t$ and exposes only a rank-matched view to each client. Let

$$\mathbf{S}_k^t = \mathcal{P}_{r_k}(\mathbf{M}^t) \tag{75}$$

denote the truncated-SVD projection used by client k. The corresponding client-side effective reference is

$$\mathbf{X}_k^t = \mathbf{W}^t + \mathbf{S}_k^t, \tag{76}$$

whereas the ideal dense reference is

$$\mathbf{X}_\star^t = \mathbf{W}^t + \mathbf{M}_\star^t. \tag{77}$$

Their discrepancy satisfies

$$\begin{aligned} \|\mathbf{X}_k^t - \mathbf{X}_\star^t\|_F &= \|\mathcal{P}_{r_k}(\mathbf{M}^t) - \mathbf{M}_\star^t\|_F \\ &\leq \underbrace{\|\mathbf{M}^t - \mathbf{M}_\star^t\|_F}_{\text{memory-estimation error}} + \underbrace{\|\mathbf{M}^t - \mathcal{P}_{r_k}(\mathbf{M}^t)\|_F}_{\text{rank-projection error}}. \end{aligned} \tag{78}$$

By the Eckart–Young–Mirsky theorem,

$$\|\mathbf{M}^t - \mathcal{P}_{r_k}(\mathbf{M}^t)\|_F^2 = \sum_{i>r_k} \sigma_i^2(\mathbf{M}^t). \tag{79}$$

We therefore define the average rank-projection error as

$$\varepsilon_{\mathrm{rank},t}^2 = \sum_{k=1}^{K} p_k \sum_{i>r_k} \sigma_i^2(\mathbf{M}^t). \tag{80}$$

Under the L-smoothness assumption, the discrepancy in the dense complementary state induces

$$\mathbb{E}\|\mathbf{b}_{\mathrm{mem}}^t\|_F^2 \le L^2 \varepsilon_{\mathrm{mem},t}^2, \tag{81}$$

while capacity-aware truncation contributes

$$\mathbb{E}\|\mathbf{b}_{\mathrm{rank}}^t\|_F^2 \le L^2 \varepsilon_{\mathrm{rank},t}^2. \tag{82}$$

Hence, the former is governed by accumulated reliability-estimation error, whereas the latter is exactly controlled by the spectral energy discarded beyond each client's feasible rank.

Importantly, Residual-Sketch introduces no additional approximation term: it is an exact factorized representation of $\mathcal{P}_{r_k}(\mathbf{M}^t)$ and affects only communication, not the optimization error analyzed above.

B.4 Local Optimization Error

Suppose each client performs E local stochastic-gradient steps with local learning rate η_{loc}. Under **Assumptions 6–8**, standard local-update analysis yields

$$\mathbb{E}\|\mathbf{b}_{\mathrm{loc}}^t\|_F^2 \le C_{\mathrm{loc}} L^2 \eta_{\mathrm{loc}}^2 E^2 (\sigma^2 + \zeta^2), \tag{83}$$

where $C_{\mathrm{loc}} > 0$ is independent of the communication round. We denote this term by $\varepsilon_{\mathrm{loc}}^2$.

Combining Eqs. 81 and 82, and the local optimization bound, the total conditional bias satisfies

$$\mathbb{E}\|\mathbf{b}^t\|_F^2 \le B_t^2, \tag{84}$$

where

$$\boxed{B_t^2 = 3L^2 \varepsilon_{\mathrm{mem},t}^2 + 3L^2 \varepsilon_{\mathrm{rank},t}^2 + 3\varepsilon_{\mathrm{loc}}^2.} \tag{85}$$

B.5 Global Convergence

We now establish the global convergence result stated in **Theorem 2**. Recall from Appendix B.1 that $\mathbf{X}^t = \mathbf{W}^t + \mathbf{M}^t$ denotes the effective dense server state and that the FedRespro update can be written as

$$\mathbf{X}^{t+1} = \mathbf{X}^t - \eta_t \widehat{\mathbf{g}}^t, \tag{86}$$

where

$$\widehat{\mathbf{g}}^t = \nabla F(\mathbf{X}^t) + \boldsymbol{\xi}^t + \mathbf{b}^t, \qquad \mathbf{b}^t = \mathbf{b}_{\mathrm{mem}}^t + \mathbf{b}_{\mathrm{rank}}^t + \mathbf{b}_{\mathrm{loc}}^t. \tag{87}$$

From Appendix B.4,

$$\mathbb{E}\|\mathbf{b}^t\|_F^2 \le B_t^2, \tag{88}$$

with

$$B_t^2 = 3L^2 \varepsilon_{\mathrm{mem},t}^2 + 3L^2 \varepsilon_{\mathrm{rank},t}^2 + 3\varepsilon_{\mathrm{loc}}^2. \tag{89}$$

For a possibly varying learning rate $\eta_t \le 1/(4L)$, FedRespro satisfies

$$\frac{\sum_{t=0}^{T-1} \eta_t \mathbb{E}\|\nabla F(\mathbf{X}^t)\|_F^2}{\sum_{t=0}^{T-1} \eta_t} \le \frac{2\left(F(\mathbf{X}^0) - F^\star\right)}{\sum_{t=0}^{T-1} \eta_t} + L\sigma^2 \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} + \frac{5}{2} \frac{\sum_{t=0}^{T-1} \eta_t B_t^2}{\sum_{t=0}^{T-1} \eta_t}. \tag{90}$$

Proof. Let $\mathbf{g}^t = \nabla F(\mathbf{X}^t)$. By the L-smoothness of F,

$$F(\mathbf{X}^{t+1}) \le F(\mathbf{X}^t) - \eta_t \left\langle \mathbf{g}^t, \widehat{\mathbf{g}}^t \right\rangle_F + \frac{L\eta_t^2}{2} \|\widehat{\mathbf{g}}^t\|_F^2. \tag{91}$$

Substituting Eq. 87 gives

$$F(\mathbf{X}^{t+1}) \le F(\mathbf{X}^t) - \eta_t \|\mathbf{g}^t\|_F^2 - \eta_t \left\langle \mathbf{g}^t, \mathbf{b}^t \right\rangle_F - \eta_t \left\langle \mathbf{g}^t, \boldsymbol{\xi}^t \right\rangle_F + \frac{L\eta_t^2}{2} \|\mathbf{g}^t + \mathbf{b}^t + \boldsymbol{\xi}^t\|_F^2. \tag{92}$$

Let $\mathbb{E}_t[\cdot]$ denote expectation conditioned on the round-t history. Under **Assumption 7**, $\mathbb{E}_t[\boldsymbol{\xi}^t] = \mathbf{0}$ and $\mathbb{E}_t\|\boldsymbol{\xi}^t\|_F^2 \le \sigma^2$. Therefore,

$$\mathbb{E}_t[F(\mathbf{X}^{t+1})] \le F(\mathbf{X}^t) - \eta_t \|\mathbf{g}^t\|_F^2 - \eta_t \left\langle \mathbf{g}^t, \mathbf{b}^t \right\rangle_F + \frac{L\eta_t^2}{2} \left(2\|\mathbf{g}^t\|_F^2 + 2\|\mathbf{b}^t\|_F^2 + \sigma^2\right), \tag{93}$$

where we used $\|\mathbf{g}^t + \mathbf{b}^t\|_F^2 \le 2\|\mathbf{g}^t\|_F^2 + 2\|\mathbf{b}^t\|_F^2$.

Using Young's inequality,

$$-\left\langle \mathbf{g}^t, \mathbf{b}^t \right\rangle_F \le \frac{1}{4}\|\mathbf{g}^t\|_F^2 + \|\mathbf{b}^t\|_F^2, \tag{94}$$

we obtain

$$\mathbb{E}_t[F(\mathbf{X}^{t+1})] \le F(\mathbf{X}^t) - \left(\frac{3\eta_t}{4} - L\eta_t^2\right) \|\mathbf{g}^t\|_F^2 + \left(\eta_t + L\eta_t^2\right) \|\mathbf{b}^t\|_F^2 + \frac{L\eta_t^2}{2}\sigma^2. \tag{95}$$

For $\eta_t \le 1/(4L)$,

$$\frac{3\eta_t}{4} - L\eta_t^2 \ge \frac{\eta_t}{2}, \qquad \eta_t + L\eta_t^2 \le \frac{5\eta_t}{4}. \tag{96}$$

Together with $\mathbb{E}\|\mathbf{b}^t\|_F^2 \le B_t^2$, this gives

$$\mathbb{E}[F(\mathbf{X}^{t+1})] \le \mathbb{E}[F(\mathbf{X}^t)] - \frac{\eta_t}{2}\mathbb{E}\|\nabla F(\mathbf{X}^t)\|_F^2 + \frac{5\eta_t}{4}B_t^2 + \frac{L\eta_t^2}{2}\sigma^2. \tag{97}$$

Summing Eq. 97 over $t = 0, \dots, T-1$ and using $F(\mathbf{X}^T) \ge F^\star$ yields

$$\frac{1}{2}\sum_{t=0}^{T-1} \eta_t \mathbb{E}\|\nabla F(\mathbf{X}^t)\|_F^2 \le F(\mathbf{X}^0) - F^\star + \frac{5}{4}\sum_{t=0}^{T-1} \eta_t B_t^2 + \frac{L\sigma^2}{2}\sum_{t=0}^{T-1} \eta_t^2. \tag{98}$$

Dividing both sides by $\frac{1}{2}\sum_{t=0}^{T-1}\eta_t$ establishes Eq. 90. For a constant learning rate $\eta_t = \eta$, this reduces exactly to the form stated in **Theorem 2**.

B.6 Convergence Rate and Error Floor

For a constant learning rate $\eta_t = \eta \le 1/(4L)$, **Theorem 2** gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\mathbf{X}^t)\|_F^2 \le \frac{2\left(F(\mathbf{X}^0) - F^\star\right)}{\eta T} + L\eta\sigma^2 + \frac{5}{2T}\sum_{t=0}^{T-1} B_t^2. \tag{99}$$

Choosing $\eta = \Theta(T^{-1/2})$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\mathbf{X}^t)\|_F^2 = \mathcal{O}(T^{-1/2}) + \mathcal{O}\left(\frac{1}{T}\sum_{t=0}^{T-1} B_t^2\right). \tag{100}$$

Using Eq. 85, the FedRespro-specific approximation term satisfies

$$B_t^2 = \mathcal{O}\left(L^2 \varepsilon_{\mathrm{mem},t}^2 + L^2 \sum_{k=1}^{K} p_k \sum_{i>r_k} \sigma_i^2(\mathbf{M}^t) + \varepsilon_{\mathrm{loc}}^2 \right). \tag{101}$$

Furthermore, Eq. 74 gives

$$\varepsilon_{\mathrm{mem},t} \le \sum_{\tau=0}^{t-1} |\lambda_\tau| \varepsilon_{\mathrm{rel},\tau}, \tag{102}$$

where, from Eq. 69,

$$\varepsilon_{\mathrm{rel},\tau}^2 \le \frac{4G_R^2}{z_{\min}^2} \left(\sum_{k \in \mathcal{H}} \sqrt{\Delta_{k,s}^\tau + \Delta_{k,p}^\tau} \right)^2. \tag{103}$$

Eqs. 101–103 make explicit the roles of the two FedRespro components. Reliability-aware complementary knowledge discovery controls the accumulated memory error through the spatial and temporal reliability estimation terms, whereas capacity-aware rank-matched projection controls the spectral-tail error discarded beyond each client's feasible rank. The local term corresponds to standard federated optimization drift.

Therefore, if

$$\frac{1}{T} \sum_{t=0}^{T-1} B_t^2 \to 0, \tag{104}$$

FedRespro attains the standard $\mathcal{O}(T^{-1/2})$ stationarity rate. More generally, if

$$\sup_T \frac{1}{T} \sum_{t=0}^{T-1} B_t^2 < \infty, \tag{105}$$

FedRespro converges to a neighborhood whose size is governed jointly by the accumulated reliability error, the residual spectral tail, and local optimization drift.

C Overhead Analysis of FedRespro

C.1 Computational Overhead

We analyze the additional computational cost introduced by FedRespro. Consider a trainable matrix $\mathbf{W} \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$ and denote $D = d_{\mathrm{out}} d_{\mathrm{in}}$. Let $\mathcal{K}_t$ be the set of participating clients at round t, $\mathcal{H}_t \subseteq \mathcal{K}_t$ the high-capacity clients, $r_{\mathcal{L}}$ the rank of the statistically supported reference subspace, and $r_{\max} = \max_{k \in \mathcal{K}_t} r_k$.

Client-side computation. FedRespro does not introduce additional forward or backward passes during local training. The rank-matched residual $\mathbf{S}_k^t$ only changes the optimization reference from which client k starts its local update. Therefore, the asymptotic local training complexity remains the same as the underlying rank-r_k local optimizer. Computing the client residual $\mathbf{G}_k^t = \Delta \mathbf{W}_{k,t} - \Delta \mathbf{W}_k^{\mathrm{ref}}$ requires only a linear-time subtraction over the trainable parameters and is negligible compared with local optimization.

Server-side reliability estimation. The server first constructs the statistical reference and extracts its leading $r_{\mathcal{L}}$ singular subspace. Using truncated SVD, this requires

$$\mathcal{O}(D r_{\mathcal{L}}) \tag{106}$$

operations. Spatial consistency is evaluated by projecting high-capacity updates onto this reference subspace. For a low-rank update $\mathbf{G}_k = \mathbf{B}_k \mathbf{A}_k$ with rank r_k, the projection can be implemented directly on its factors with complexity

$$\mathcal{O}\big((d_{\mathrm{out}} + d_{\mathrm{in}}) r_{\mathcal{L}} r_k + r_{\mathcal{L}}^2 r_k \big), \tag{107}$$

rather than materializing an additional dense matrix. Cross-round persistence only requires an inner product between the current residual and the stored residual state and therefore contributes at most $\mathcal{O}(D)$ per high-capacity client.

Cross-round preservation and rank-matched projection. Updating the dense residual memory

$$\mathbf{M}^t = \mathbf{M}^{t-1} + \lambda_t \mathbf{R}^t_{\text{rel}} \tag{108}$$

requires $\mathcal{O}(D)$ operations. Importantly, FedRespro only needs the leading $r_{\max}$ singular directions of $\mathbf{M}^t$. Hence, a truncated SVD requires

$$\mathcal{O}(Dr_{\max}), \tag{109}$$

after which the reference for each client is obtained by simply selecting the first r_k components, without repeating the decomposition.

Overall, excluding the local training cost shared with the underlying FL optimizer, the additional server-side complexity per round is

$$\mathcal{C}_{\text{FedRespro}} = \mathcal{O}\Bigg(D(r_{\mathcal{L}} + r_{\max} + |\mathcal{H}_t|) + \sum_{k\in\mathcal{H}_t} (d_{\text{out}} + d_{\text{in}}) r_{\mathcal{L}} r_k \Bigg), \tag{110}$$

where lower-order terms are omitted. Since typically $r_{\mathcal{L}}, r_k \ll \min(d_{\text{out}}, d_{\text{in}})$, the dominant additional computation is shifted to the server and does not increase the computational capacity required at constrained clients.

FedRespro additionally stores the complete residual state $\mathbf{M}^t$ on the server, requiring $\mathcal{O}(D)$ server memory, whereas each client only maintains its rank-r_k representation with $\mathcal{O}((d_{\text{out}} + d_{\text{in}})r_k)$ parameters. For multi-layer models, the above costs extend linearly over all adapted layers.

C.2 Communication Overhead

We analyze the communication overhead of FedRespro under the Residual-Sketch implementation with a separate residual channel. This analysis is distinct from the adapter-only measurements in Appendix H. The key design is to decouple *server-side knowledge preservation* from *client-side knowledge transmission*: the accumulated residual $\mathbf{M}^t$ is maintained as an exact dense state at the server, whereas only a rank-matched low-rank sketch is transmitted to each client.

For a trainable matrix $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, define the number of scalars required by a rank-r representation as

$$q(r) = r(d_{\text{out}} + d_{\text{in}}). \tag{111}$$

Client k uploads its rank-r_k local update in factorized form, requiring

$$C_{k,\text{up}} = q(r_k) \tag{112}$$

scalars, which is identical to standard heterogeneous low-rank federated training.

Residual-Sketch downlink. FedRespro does not transmit the dense residual $\mathbf{M}^t \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Instead, for client k, the server determines the sketch rank as

$$s_k = \min\left(s_{\max}, \lceil \mu r_k \rceil\right), \tag{113}$$

where $\mu = 1$ corresponds to the rank-matched configuration used by FedRespro. Given

$$\mathbf{M}^t = \mathbf{U}^t \mathbf{\Sigma}^t (\mathbf{V}^t)^\top, \tag{114}$$

the transmitted residual sketch is

$$\mathcal{P}_{s_k}(\mathbf{M}^t) = \mathbf{U}^t_{1:s_k} \mathbf{\Sigma}^t_{1:s_k} (\mathbf{V}^t_{1:s_k})^\top. \tag{115}$$

It is factorized as

$$\mathcal{P}_{s_k}(\mathbf{M}^t) = \underbrace{\mathbf{U}^t_{1:s_k} (\mathbf{\Sigma}^t_{1:s_k})^{1/2}}_{\mathbf{B}^t_{k,\text{sk}}} \underbrace{(\mathbf{\Sigma}^t_{1:s_k})^{1/2} (\mathbf{V}^t_{1:s_k})^\top}_{\mathbf{A}^t_{k,\text{sk}}}, \tag{116}$$

and therefore requires only

$$C_{k,\text{sk}} = q(s_k) = s_k(d_{\text{out}} + d_{\text{in}}) \tag{117}$$

scalars.

The fixed model reference is transmitted once and can be cached locally; thus, it does not contribute to the steady-state per-round overhead. Moreover, statistical support, spatial consistency, cross-round persistence, and reliability weights are all computed at the server and require only negligible scalar metadata.

Per-round communication. Let $\mathcal{K}_t$ denote the participating clients in round t. A conventional rank-matched low-rank protocol requires approximately

$$C^t_{\text{base}} = 2 \sum_{k \in \mathcal{K}_t} q(r_k) \tag{118}$$

scalars for low-rank uplink and downlink. Residual-Sketch adds one rank-s_k residual view, giving

$$C^t_{\text{FedRespro}} = \sum_{k \in \mathcal{K}_t} [2q(r_k) + q(s_k)] + \mathcal{O}(|\mathcal{K}_t|). \tag{119}$$

Hence, the FedRespro-specific additional communication is

$$\Delta C^t_{\text{FedRespro}} = \sum_{k \in \mathcal{K}_t} q(s_k) + \mathcal{O}(|\mathcal{K}_t|). \tag{120}$$

For the default rank-matched setting $s_k = r_k$,

$$\frac{C^t_{\text{FedRespro}}}{C^t_{\text{base}}} \approx \frac{3}{2}. \tag{121}$$

Thus, Residual-Sketch introduces a moderate communication increase over standard low-rank federated training in exchange for transmitting cross-round complementary residual knowledge.

Reduction over dense residual transmission. The important communication benefit arises relative to transmitting the complete accumulated residual. A dense residual requires

$$D = d_{\text{out}} d_{\text{in}} \tag{122}$$

scalars per client, whereas Residual-Sketch requires only $q(s_k)$. The communication ratio of the residual channel is therefore

$$\rho_k = \frac{q(s_k)}{D} = \frac{s_k(d_{\text{out}} + d_{\text{in}})}{d_{\text{out}} d_{\text{in}}}. \tag{123}$$

For a square $d \times d$ matrix,

$$\rho_k = \frac{2s_k}{d}, \tag{124}$$

which becomes small when $s_k \ll d$.

Therefore, FedRespro does not reduce communication by compressing the server-side knowledge state itself. Instead, it preserves the complete dense residual at the server while exposing only capacity-matched low-rank sketches to clients. This reduces the residual transmission from $\mathcal{O}(d_{\text{out}} d_{\text{in}})$ to $\mathcal{O}(s_k(d_{\text{out}} + d_{\text{in}}))$, without recursively introducing sketching errors into the server-side residual trajectory.

D Experimental Design and Reporting Conventions

D.1 Datasets, Models, and Optimization

Qwen2-0.5B/1.5B/7B are fine-tuned on the same 170,420-example Commonsense170K training corpus. Complete prompt–response sequences are retained with a 2,048-token training limit. Downstream generation uses eight fixed 500-example task subsets, a 512-token input limit, at most 32 generated tokens, and the same answer-extraction protocol. The sampling seed is 20260829. The reported score is the unweighted mean of eight task accuracies. All accepted evaluations contain zero truncated inputs. R4 and R8 denote the corresponding client deployment views; the server is evaluated before client-specific dispatch.

BERT-base-uncased uses a trainable classifier and query/value adapters. CoLA contains 8,551 training-pool examples and all 1,043 labeled validation examples; we report MCC multiplied by 100. DBpedia-14 uses a fixed 30,000-example training subset and a stratified 3,000-example test subset; we report accuracy in percent. Client-local validation reserves 8%/5% of the CoLA/DBpedia partitions. The Qwen local validation fraction is 5%.

Table A1: Shared settings for Qwen2 and BERT experiments. α_D is the Dirichlet label-skew concentration and σ_n controls lognormal quantity skew; rank proportions specify the fraction of clients at each listed rank.

Setting	Qwen2	CoLA	DBpedia
Clients / selected each round	40 / 6	20 / 5	20 / 5
Rounds / local steps	40 / 10	20 / 20	15 / 20
Trainable ranks	4, 8	2, 4, 8	2, 4, 8
Rank proportions	75/25%	35/35/30%	35/35/30%
Learning rate	10^{-4}	2×10^{-4}	2×10^{-4}
Effective batch size	4	32	32
Maximum training tokens	2,048	128	192
α_D / σ_n	0.3 / 1.2	0.5 / 0.7	0.5 / 1.0
Optimizer / precision	AdamW / bf16	AdamW / bf16	AdamW / bf16

The low-rank adapter scale satisfies $\alpha/r = 2$, with zero adapter dropout. Adapters target query and value projections in both model families. Local optimization uses a cosine schedule, zero weight decay, and gradient clipping at norm 1. Qwen uses microbatches of 2/2/1 with accumulation steps 2/2/4 at 0.5B/1.5B/7B. Larger sample counts are assigned smaller trainable ranks. Qwen uses gradient checkpointing without quantization. Native aggregation rules and parameterizations are retained for the five external baselines. Auxiliary classifier parameters are included in the communication accounting; traffic is measured from the tensors transmitted by each implementation.

D.2 Training Seeds and Reporting Units

Unless a table states otherwise, uncertainty is the sample standard deviation across training seeds on one fixed data partition. For paired studies, differences are computed within each seed before summarization. Each training seed contributes one result.

Table A2: Experimental studies and their reporting units. Standard deviations are computed over the listed unit for each study.

Study	Controlled unit	Reported outcome
Qwen downstream	Training seed / model	Eight-task and per-task accuracy
Encoder downstream	Training seed / dataset	MCC or accuracy
Communication	Training run	Upload and download bytes
Hyperparameter sweep	Fixed seed / coefficient	Eight-task server accuracy
Residual interventions	Paired training seed	Absolute scores and paired differences

Plot conventions. Markers show measured conditions or rounds, and lines connect adjacent observations. Shaded bands and error bars denote sample SD where stated. Orange identifies the FedRespro configuration specified in each caption.

Downstream panel. The eight tasks are BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC-Easy, ARC-Challenge, and OpenBookQA. At 0.5B, internal and external configurations share the data partition and use different client-sampling and initialization sub-seeds.

E TASK-LEVEL DOWNSTREAM RESULTS

Per-task accuracy uses the fixed 500-example subsets and 40-round protocol in Appendix D. Entries show mean and sample SD over three seeds. Each reported client view uses its corresponding server checkpoint. Each model uses the same FedRespro configuration as the main comparison.

E.1 QWEN2-0.5B

Table A3: Qwen2-0.5B rank-4 client accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$52.5_{\pm 12.3}$	$53.7_{\pm 1.1}$	$39.9_{\pm 1.4}$	$26.5_{\pm 0.2}$	$47.6_{\pm 0.4}$	$29.4_{\pm 2.3}$	$28.9_{\pm 0.9}$	$23.1_{\pm 2.3}$	$37.7_{\pm 2.2}$
FlexLoRA	$61.1_{\pm 5.6}$	$53.6_{\pm 2.8}$	$41.1_{\pm 1.5}$	$25.6_{\pm 1.0}$	$48.2_{\pm 1.1}$	$31.9_{\pm 3.2}$	$29.7_{\pm 1.0}$	$26.4_{\pm 2.2}$	$39.7_{\pm 1.0}$
FLoRA	$61.9_{\pm 4.4}$	$56.8_{\pm 1.5}$	$38.8_{\pm 0.9}$	$26.5_{\pm 0.8}$	$47.7_{\pm 0.7}$	$32.1_{\pm 2.4}$	$28.4_{\pm 0.4}$	$26.1_{\pm 1.1}$	$39.8_{\pm 0.5}$
RB-LoRA	$57.1_{\pm 10.5}$	$53.6_{\pm 3.2}$	$41.2_{\pm 1.3}$	$25.8_{\pm 0.5}$	$47.8_{\pm 0.5}$	$31.1_{\pm 2.7}$	$30.1_{\pm 0.5}$	$24.6_{\pm 2.9}$	$38.9_{\pm 2.1}$
FSLoRA	$58.0_{\pm 8.1}$	$57.0_{\pm 0.9}$	$39.9_{\pm 0.8}$	$25.5_{\pm 1.0}$	$47.9_{\pm 1.4}$	$35.7_{\pm 1.9}$	$29.1_{\pm 1.1}$	$29.0_{\pm 2.3}$	$40.3_{\pm 0.8}$
FedRespro	$59.7_{\pm 8.5}$	$53.0_{\pm 0.3}$	$39.7_{\pm 3.8}$	$24.9_{\pm 1.6}$	$49.3_{\pm 1.5}$	$35.9_{\pm 9.8}$	$27.3_{\pm 6.3}$	$30.3_{\pm 4.7}$	$40.0_{\pm 3.5}$

Table A4: Qwen2-0.5B rank-8 client accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$59.5_{\pm 5.6}$	$54.4_{\pm 2.3}$	$38.9_{\pm 1.5}$	$26.7_{\pm 0.8}$	$48.1_{\pm 0.1}$	$29.1_{\pm 2.8}$	$27.9_{\pm 1.5}$	$24.0_{\pm 1.6}$	$38.6_{\pm 1.4}$
FlexLoRA	$59.5_{\pm 8.1}$	$54.1_{\pm 2.5}$	$39.9_{\pm 0.3}$	$26.3_{\pm 1.7}$	$48.3_{\pm 1.0}$	$30.9_{\pm 2.4}$	$29.3_{\pm 1.2}$	$25.8_{\pm 2.3}$	$39.3_{\pm 1.4}$
FLoRA	$61.9_{\pm 4.4}$	$56.8_{\pm 1.5}$	$38.8_{\pm 0.9}$	$26.5_{\pm 0.8}$	$47.7_{\pm 0.7}$	$32.1_{\pm 2.4}$	$28.4_{\pm 0.4}$	$26.1_{\pm 1.1}$	$39.8_{\pm 0.5}$
RB-LoRA	$47.0_{\pm 15.0}$	$52.6_{\pm 1.1}$	$39.9_{\pm 1.4}$	$25.9_{\pm 0.9}$	$49.6_{\pm 2.1}$	$30.9_{\pm 3.4}$	$28.9_{\pm 1.4}$	$25.1_{\pm 2.4}$	$37.5_{\pm 1.8}$
FSLoRA	$54.6_{\pm 8.7}$	$56.1_{\pm 1.3}$	$40.5_{\pm 1.1}$	$25.8_{\pm 0.8}$	$48.7_{\pm 1.5}$	$37.4_{\pm 1.5}$	$30.7_{\pm 2.9}$	$29.5_{\pm 1.7}$	$40.4_{\pm 0.9}$
FedRespro	$54.9_{\pm 16.9}$	$53.0_{\pm 0.3}$	$39.9_{\pm 4.4}$	$24.9_{\pm 1.6}$	$49.0_{\pm 2.5}$	$34.6_{\pm 6.5}$	$26.9_{\pm 5.0}$	$30.0_{\pm 3.8}$	$39.1_{\pm 3.8}$

Table A5: Qwen2-0.5B server accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$60.1_{\pm 5.7}$	$55.2_{\pm 1.9}$	$39.5_{\pm 0.9}$	$25.9_{\pm 0.7}$	$48.2_{\pm 0.7}$	$30.1_{\pm 3.2}$	$28.3_{\pm 1.5}$	$24.5_{\pm 2.2}$	$39.0_{\pm 1.5}$
FlexLoRA	$60.3_{\pm 6.9}$	$54.3_{\pm 4.3}$	$40.1_{\pm 1.6}$	$25.8_{\pm 0.4}$	$48.7_{\pm 1.3}$	$31.8_{\pm 3.5}$	$29.5_{\pm 0.2}$	$25.5_{\pm 2.8}$	$39.5_{\pm 1.7}$
FLoRA	$61.3_{\pm 5.4}$	$56.0_{\pm 0.3}$	$39.3_{\pm 0.9}$	$26.7_{\pm 0.6}$	$48.2_{\pm 0.2}$	$32.9_{\pm 1.2}$	$28.0_{\pm 2.0}$	$25.7_{\pm 0.8}$	$39.8_{\pm 0.9}$
RB-LoRA	$47.3_{\pm 14.7}$	$52.9_{\pm 1.2}$	$40.3_{\pm 2.0}$	$24.7_{\pm 1.5}$	$48.9_{\pm 1.7}$	$31.5_{\pm 3.9}$	$29.9_{\pm 1.3}$	$25.5_{\pm 3.1}$	$37.6_{\pm 1.8}$
FSLoRA	$55.5_{\pm 7.9}$	$55.3_{\pm 0.8}$	$41.1_{\pm 0.6}$	$25.6_{\pm 0.2}$	$47.4_{\pm 0.5}$	$37.2_{\pm 2.0}$	$30.5_{\pm 3.1}$	$30.4_{\pm 0.7}$	$40.4_{\pm 0.8}$
FedRespro	$64.6_{\pm 0.0}$	$55.0_{\pm 2.0}$	$41.7_{\pm 6.1}$	$27.1_{\pm 2.2}$	$48.1_{\pm 0.1}$	$35.3_{\pm 9.1}$	$27.5_{\pm 5.1}$	$30.7_{\pm 5.3}$	$41.2_{\pm 3.6}$

E.2 QWEN2-1.5B

Table A6: Qwen2-1.5B rank-4 client accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$62.7_{\pm 0.4}$	$67.8_{\pm 0.7}$	$59.3_{\pm 1.5}$	$35.7_{\pm 3.9}$	$47.7_{\pm 1.4}$	$74.6_{\pm 0.4}$	$55.3_{\pm 0.9}$	$51.5_{\pm 1.4}$	$56.8_{\pm 0.6}$
FlexLoRA	$63.3_{\pm 1.1}$	$68.3_{\pm 0.8}$	$61.0_{\pm 0.4}$	$40.0_{\pm 6.8}$	$48.6_{\pm 0.5}$	$76.9_{\pm 0.8}$	$56.6_{\pm 0.9}$	$53.8_{\pm 0.6}$	$58.6_{\pm 0.6}$
FLoRA	$61.2_{\pm 1.2}$	$66.5_{\pm 0.5}$	$56.2_{\pm 2.3}$	$35.5_{\pm 1.6}$	$47.7_{\pm 1.7}$	$71.5_{\pm 1.7}$	$52.3_{\pm 0.6}$	$48.8_{\pm 1.6}$	$55.0_{\pm 0.8}$
RB-LoRA	$62.9_{\pm 1.4}$	$68.5_{\pm 0.3}$	$60.5_{\pm 0.6}$	$39.9_{\pm 6.3}$	$48.2_{\pm 0.7}$	$77.3_{\pm 1.0}$	$56.3_{\pm 0.1}$	$53.7_{\pm 1.0}$	$58.4_{\pm 0.5}$
FSLoRA	$62.9_{\pm 0.4}$	$68.9_{\pm 0.3}$	$60.9_{\pm 0.2}$	$40.8_{\pm 4.5}$	$47.4_{\pm 1.9}$	$76.5_{\pm 1.4}$	$54.8_{\pm 0.9}$	$53.7_{\pm 0.8}$	$58.2_{\pm 1.1}$
FedRespro	$56.1_{\pm 6.3}$	$72.1_{\pm 1.2}$	$65.1_{\pm 1.7}$	$54.3_{\pm 7.0}$	$50.9_{\pm 2.4}$	$79.4_{\pm 0.7}$	$57.3_{\pm 0.9}$	$58.8_{\pm 3.2}$	$61.7_{\pm 1.3}$

Table A7: Qwen2-1.5B rank-8 client accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$62.3_{\pm 0.3}$	$67.5_{\pm 0.7}$	$59.4_{\pm 1.6}$	$35.1_{\pm 3.4}$	$47.2_{\pm 0.3}$	$74.9_{\pm 0.2}$	$54.3_{\pm 1.3}$	$51.7_{\pm 1.1}$	$56.5_{\pm 0.3}$
FlexLoRA	$63.4_{\pm 1.0}$	$68.3_{\pm 1.4}$	$60.6_{\pm 0.5}$	$39.9_{\pm 6.9}$	$48.7_{\pm 1.1}$	$76.8_{\pm 1.6}$	$56.2_{\pm 0.7}$	$53.3_{\pm 0.5}$	$58.4_{\pm 0.8}$
FLoRA	$61.2_{\pm 1.2}$	$66.5_{\pm 0.5}$	$56.2_{\pm 2.3}$	$35.5_{\pm 1.6}$	$47.7_{\pm 1.7}$	$71.5_{\pm 1.7}$	$52.3_{\pm 0.6}$	$48.8_{\pm 1.6}$	$55.0_{\pm 0.8}$
RB-LoRA	$63.0_{\pm 1.2}$	$68.6_{\pm 0.6}$	$60.6_{\pm 0.7}$	$40.5_{\pm 6.4}$	$48.3_{\pm 0.5}$	$76.9_{\pm 0.8}$	$56.0_{\pm 1.1}$	$53.3_{\pm 0.6}$	$58.4_{\pm 0.7}$
FSLoRA	$61.9_{\pm 0.5}$	$69.2_{\pm 0.4}$	$59.6_{\pm 0.0}$	$39.9_{\pm 4.0}$	$47.2_{\pm 0.9}$	$77.3_{\pm 0.6}$	$54.7_{\pm 1.1}$	$53.8_{\pm 1.2}$	$58.0_{\pm 0.5}$
FedRespro	$52.7_{\pm 9.7}$	$72.8_{\pm 2.0}$	$65.5_{\pm 0.8}$	$53.2_{\pm 5.4}$	$50.5_{\pm 2.8}$	$78.7_{\pm 0.9}$	$58.6_{\pm 1.2}$	$60.0_{\pm 2.4}$	$61.5_{\pm 1.2}$

Table A8: Qwen2-1.5B server accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$62.5_{\pm 0.3}$	$67.5_{\pm 0.5}$	$58.7_{\pm 1.9}$	$34.9_{\pm 3.3}$	$47.7_{\pm 0.5}$	$75.1_{\pm 0.4}$	$54.8_{\pm 0.9}$	$51.5_{\pm 0.6}$	$56.6_{\pm 0.2}$
FlexLoRA	$62.9_{\pm 1.3}$	$68.7_{\pm 1.0}$	$61.1_{\pm 0.5}$	$39.8_{\pm 6.6}$	$48.7_{\pm 0.4}$	$76.4_{\pm 0.7}$	$56.3_{\pm 0.6}$	$53.7_{\pm 1.0}$	$58.4_{\pm 0.8}$
FLoRA	$60.7_{\pm 1.2}$	$66.8_{\pm 0.7}$	$55.9_{\pm 2.2}$	$35.8_{\pm 2.9}$	$48.0_{\pm 0.9}$	$72.0_{\pm 2.5}$	$51.5_{\pm 0.2}$	$49.0_{\pm 1.3}$	$55.0_{\pm 0.4}$
RB-LoRA	$62.8_{\pm 1.4}$	$68.4_{\pm 0.7}$	$60.7_{\pm 0.2}$	$40.6_{\pm 6.3}$	$48.2_{\pm 0.8}$	$77.1_{\pm 0.4}$	$55.9_{\pm 0.9}$	$53.4_{\pm 0.6}$	$58.4_{\pm 0.5}$
FSLoRA	$61.8_{\pm 0.5}$	$69.4_{\pm 0.4}$	$60.5_{\pm 0.1}$	$39.7_{\pm 4.3}$	$47.0_{\pm 0.9}$	$77.3_{\pm 1.2}$	$54.9_{\pm 0.5}$	$53.7_{\pm 1.8}$	$58.0_{\pm 0.5}$
FedRespro	$61.9_{\pm 2.4}$	$72.5_{\pm 0.6}$	$65.3_{\pm 2.4}$	$58.7_{\pm 3.9}$	$50.5_{\pm 2.0}$	$79.9_{\pm 0.6}$	$57.7_{\pm 0.4}$	$59.3_{\pm 2.2}$	$63.2_{\pm 1.2}$

E.3 QWEN2-7B

Table A9: Qwen2-7B rank-4 client accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$70.6_{\pm 0.5}$	$80.2_{\pm 0.2}$	$75.4_{\pm 0.2}$	$80.7_{\pm 0.1}$	$61.5_{\pm 0.6}$	$91.1_{\pm 0.6}$	$81.3_{\pm 0.5}$	$75.8_{\pm 0.3}$	$77.1_{\pm 0.1}$
FlexLoRA	$70.9_{\pm 0.5}$	$80.9_{\pm 1.0}$	$75.5_{\pm 1.0}$	$82.5_{\pm 0.1}$	$62.8_{\pm 0.9}$	$92.7_{\pm 0.2}$	$81.7_{\pm 0.6}$	$76.6_{\pm 0.2}$	$78.0_{\pm 0.1}$
FLoRA	$66.2_{\pm 1.7}$	$78.2_{\pm 0.2}$	$74.3_{\pm 0.5}$	$79.4_{\pm 0.3}$	$59.2_{\pm 1.4}$	$88.1_{\pm 0.6}$	$79.9_{\pm 0.2}$	$73.0_{\pm 0.3}$	$74.8_{\pm 0.3}$
RB-LoRA	$71.0_{\pm 0.5}$	$81.9_{\pm 0.8}$	$75.7_{\pm 0.8}$	$83.5_{\pm 0.8}$	$62.9_{\pm 0.7}$	$92.5_{\pm 0.2}$	$82.0_{\pm 0.7}$	$77.2_{\pm 0.3}$	$78.3_{\pm 0.0}$
FSLoRA	$70.3_{\pm 1.3}$	$82.7_{\pm 0.6}$	$75.6_{\pm 1.6}$	$83.7_{\pm 1.5}$	$62.5_{\pm 1.6}$	$92.5_{\pm 0.7}$	$83.1_{\pm 0.3}$	$78.5_{\pm 0.7}$	$78.6_{\pm 0.4}$
FedRespro	$69.8_{\pm 0.2}$	$83.3_{\pm 1.3}$	$75.6_{\pm 0.9}$	$85.1_{\pm 0.9}$	$65.2_{\pm 1.1}$	$92.9_{\pm 0.6}$	$82.1_{\pm 0.7}$	$79.0_{\pm 0.7}$	$79.1_{\pm 0.2}$

Table A10: Qwen2-7B rank-8 client accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$70.4_{\pm 0.4}$	$80.4_{\pm 0.2}$	$75.6_{\pm 0.2}$	$80.9_{\pm 0.1}$	$60.7_{\pm 0.8}$	$91.1_{\pm 0.4}$	$81.5_{\pm 1.0}$	$76.0_{\pm 0.9}$	$77.1_{\pm 0.2}$
FlexLoRA	$70.9_{\pm 1.0}$	$81.1_{\pm 0.9}$	$75.7_{\pm 0.8}$	$82.4_{\pm 0.4}$	$62.8_{\pm 0.7}$	$92.7_{\pm 0.3}$	$81.6_{\pm 0.5}$	$76.9_{\pm 0.2}$	$78.0_{\pm 0.2}$
FLoRA	$66.2_{\pm 1.7}$	$78.2_{\pm 0.2}$	$74.3_{\pm 0.5}$	$79.4_{\pm 0.3}$	$59.2_{\pm 1.4}$	$88.1_{\pm 0.6}$	$79.9_{\pm 0.2}$	$73.0_{\pm 0.3}$	$74.8_{\pm 0.3}$
RB-LoRA	$71.2_{\pm 0.2}$	$81.9_{\pm 0.5}$	$75.9_{\pm 0.7}$	$83.4_{\pm 0.9}$	$62.3_{\pm 1.6}$	$92.3_{\pm 0.5}$	$82.2_{\pm 0.4}$	$77.5_{\pm 0.3}$	$78.3_{\pm 0.1}$
FSLoRA	$70.8_{\pm 1.1}$	$83.2_{\pm 0.4}$	$76.0_{\pm 1.2}$	$83.9_{\pm 1.9}$	$63.3_{\pm 1.4}$	$92.7_{\pm 0.2}$	$82.9_{\pm 0.6}$	$78.5_{\pm 0.8}$	$78.9_{\pm 0.4}$
FedRespro	$70.1_{\pm 0.5}$	$84.1_{\pm 0.9}$	$74.9_{\pm 1.1}$	$85.6_{\pm 0.8}$	$65.3_{\pm 1.9}$	$93.3_{\pm 0.2}$	$82.9_{\pm 0.4}$	$79.9_{\pm 0.6}$	$79.5_{\pm 0.2}$

Table A11: Qwen2-7B server accuracy (%) by task. Values are means with sample SD over three seeds.

Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
HetLoRA	$70.8_{\pm 0.3}$	$80.5_{\pm 0.2}$	$75.4_{\pm 0.5}$	$80.8_{\pm 0.3}$	$60.8_{\pm 0.4}$	$91.0_{\pm 0.6}$	$81.3_{\pm 0.8}$	$75.7_{\pm 0.6}$	$77.0_{\pm 0.0}$
FlexLoRA	$70.9_{\pm 0.5}$	$80.9_{\pm 0.9}$	$75.6_{\pm 0.7}$	$82.5_{\pm 0.3}$	$63.0_{\pm 0.6}$	$92.3_{\pm 0.6}$	$81.7_{\pm 0.2}$	$76.5_{\pm 0.4}$	$77.9_{\pm 0.1}$
FLoRA	$66.6_{\pm 2.0}$	$78.1_{\pm 0.6}$	$74.3_{\pm 0.3}$	$79.3_{\pm 0.6}$	$58.7_{\pm 1.4}$	$88.2_{\pm 1.1}$	$79.7_{\pm 0.1}$	$73.5_{\pm 0.4}$	$74.8_{\pm 0.2}$
RB-LoRA	$71.3_{\pm 0.1}$	$81.7_{\pm 0.6}$	$75.5_{\pm 0.8}$	$83.7_{\pm 0.6}$	$62.2_{\pm 1.4}$	$92.2_{\pm 0.5}$	$82.2_{\pm 0.8}$	$76.8_{\pm 0.3}$	$78.2_{\pm 0.2}$
FSLoRA	$70.7_{\pm 1.1}$	$83.4_{\pm 0.5}$	$75.9_{\pm 1.3}$	$84.2_{\pm 1.7}$	$63.4_{\pm 1.5}$	$92.6_{\pm 0.2}$	$82.8_{\pm 0.7}$	$78.3_{\pm 0.8}$	$78.9_{\pm 0.4}$
FedRespro	$70.0_{\pm 0.3}$	$84.5_{\pm 0.8}$	$75.7_{\pm 1.0}$	$85.7_{\pm 0.8}$	$66.6_{\pm 1.1}$	$93.3_{\pm 0.1}$	$82.9_{\pm 0.6}$	$80.2_{\pm 0.7}$	$79.8_{\pm 0.1}$

E.4 Task and Deployment-View Differences

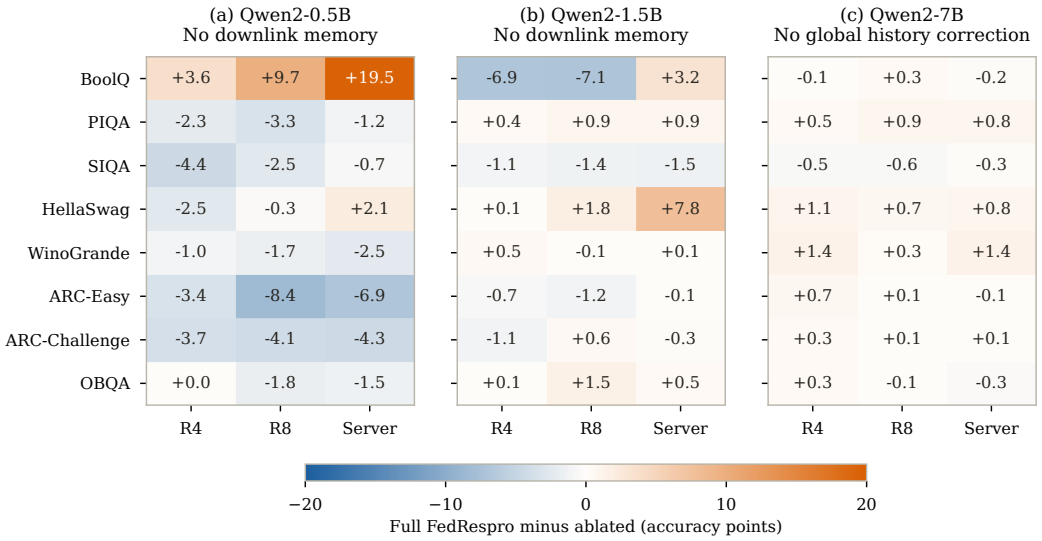

Figure A1: Task-level differences for the selected main configurations, averaged over three seeds. Each cell is full minus ablated accuracy in points. Ablations remove downlink memory at 0.5B/1.5B and global history correction at 7B; columns show R4, R8, and server views on a common color scale.

F Encoder Protocol and Additional Results

F.1 Fixed-Configuration Main Comparison

Both encoder tasks use BERT-base-uncased with a trainable classifier and query/value adapters. CoLA uses 8,551 training-pool examples and all 1,043 labeled validation examples; DBpedia-14 uses a fixed 30,000-example training subset and a stratified 3,000-example test subset. Local validation reserves 8%/5% of each client partition for CoLA/DBpedia. Label proportions follow a Dirichlet distribution with concentration 0.5; lognormal quantity-skew parameters are 0.7/1.0. Input limits are 128/192 tokens. Clients have ranks 2/4/8 in proportions 35/35/30%, inversely assigned to sample counts. Five of 20 clients participate per round. AdamW uses learning rate 2×10^{-4}, 20 local steps, batch size 32, and 20/15 rounds. Each method uses its native aggregation rule and adapter parameterization.

Table A12: BERT-base-uncased results with full uncertainty: CoLA MCC $\times 100$ and DBpedia accuracy (%). Values are mean $\pm$ sample SD over three seeds. "failed" denotes failure to converge.

Method	CoLA	DBpedia
HetLoRA	failed	15.58 ± 0.94
FlexLoRA	6.17 ± 1.38	19.89 ± 3.34
FLoRA	0.60 ± 1.05	14.12 ± 0.87
RB-LoRA	19.57 ± 11.28	26.83 ± 2.81
FSLoRA	11.55 ± 7.12	19.32 ± 1.22
FedRespro	43.00 ± 0.83	43.64 ± 2.51

G Component Ablation Results

G.1 Effects of Residual Components

Table A13 and Figure A2 show how residual components affect server and client performance. Downlink memory improves the server means by 0.55 and 1.33 percentage points at Qwen2-0.5B and 1.5B. At 7B, global correction improves all three views, with gains of 0.46, 0.21, and 0.28 points for R4, R8, and the server. CoLA shows a larger dependence on residual history: cumulative residuals improve MCC by 27.65 points over round-only residuals and reduce the standard deviation from 4.13 to 0.83.

The additional variants in Table 3 reinforce these component benefits. Without global correction, server accuracy falls to 40.56% at 0.5B and 59.07% at 1.5B, compared with 41.24% and 63.23%

for full FedRespro. Removing downlink memory reduces 7B accuracy from 79.84% to 78.37% and fails on CoLA.

Table A13: Matched ablations: mean ± sample SD over seeds 0/1/2. Full is FedRespro. Ablated denotes *w/o downlink memory* for Qwen2-0.5B/1.5B, *w/o global correction* for Qwen2-7B, and *round-only residual* for CoLA. Qwen uses accuracy (%); CoLA uses MCC ×100. Δ is full minus ablated.

Model/task	View	Full	Ablated	Δ: full−ablated
Qwen2-0.5B	R4	40.02 ± 3.53	41.74 ± 3.07	-1.73 ± 2.46
Qwen2-0.5B	R8	39.14 ± 3.84	40.69 ± 3.22	-1.55 ± 3.62
Qwen2-0.5B	Server	41.24 ± 3.63	40.69 ± 3.11	$+0.55 \pm 4.51$
Qwen2-1.5B	R4	61.74 ± 1.29	62.84 ± 0.51	-1.10 ± 0.78
Qwen2-1.5B	R8	61.50 ± 1.21	62.13 ± 0.68	-0.63 ± 0.54
Qwen2-1.5B	Server	63.23 ± 1.18	61.91 ± 0.90	$+1.33 \pm 0.78$
Qwen2-7B	R4	79.13 ± 0.16	78.67 ± 0.08	$+0.46 \pm 0.14$
Qwen2-7B	R8	79.51 ± 0.18	79.30 ± 0.18	$+0.21 \pm 0.19$
Qwen2-7B	Server	79.84 ± 0.07	79.57 ± 0.17	$+0.28 \pm 0.23$
CoLA	Server MCC	43.00 ± 0.83	15.35 ± 4.13	$+27.65 \pm 3.53$

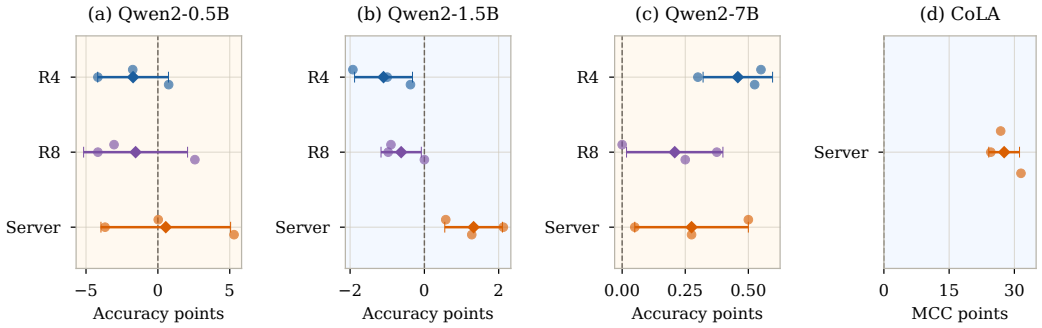

Figure A2: FedRespro minus the ablated variant: *w/o downlink memory* at Qwen2-0.5B/1.5B, *w/o global correction* at Qwen2-7B, and *round-only residual* for CoLA. Circles show individual seeds; diamonds and error bars show mean ± sample SD. Qwen panels report accuracy-point differences for R4, R8, and server views; CoLA reports server MCC points.

G.2 DBpedia Three-Seed Results

Table A14 shows that cumulative residual history improves DBpedia accuracy in every seed over round-only residuals, raising the mean from 14.07% to 43.64%. Downlink memory mainly improves consistency: the mean changes from 43.57% to 43.64%, while the standard deviation falls from 11.56 to 2.51. Full FedRespro combines the highest mean accuracy with the lowest variation across these three variants.

Table A14: DBpedia server accuracy (%). Summaries are mean ± sample SD over three seeds, computed before rounding the displayed seed scores.

Configuration	Seed 0	Seed 1	Seed 2	Mean ± SD
FedRespro	46.53	42.07	42.33	43.64 ± 2.51
w/o downlink memory	55.77	32.77	42.17	43.57 ± 11.56
Round-only residual	12.77	19.03	10.40	14.07 ± 4.46

G.3 Individual Matched Seeds

Table A15: Individual matched-seed scores. Full is FedRespro. Ablated denotes *w/o downlink memory* for Qwen2-0.5B/1.5B, *w/o global correction* for Qwen2-7B, and *round-only residual* for CoLA. Qwen uses eight-task accuracy (%); CoLA uses MCC $\times 100$. Δ is full minus ablated at the same seed and checkpoint rule.

Model/task	View	Seed	Full	Ablated	Δ
Qwen2-0.5B	R4	0	43.575	42.825	+0.750
Qwen2-0.5B	R4	1	39.950	44.125	-4.175
Qwen2-0.5B	R4	2	36.525	38.275	-1.750
Qwen2-0.5B	R8	0	42.425	39.850	+2.575
Qwen2-0.5B	R8	1	40.075	44.250	-4.175
Qwen2-0.5B	R8	2	34.925	37.975	-3.050
Qwen2-0.5B	Server	0	45.200	39.900	+5.300
Qwen2-0.5B	Server	1	40.450	44.125	-3.675
Qwen2-0.5B	Server	2	38.075	38.050	+0.025
Qwen2-1.5B	R4	0	62.975	63.350	-0.375
Qwen2-1.5B	R4	1	61.850	62.850	-1.000
Qwen2-1.5B	R4	2	60.400	62.325	-1.925
Qwen2-1.5B	R8	0	62.900	62.900	0.000
Qwen2-1.5B	R8	1	60.850	61.825	-0.975
Qwen2-1.5B	R8	2	60.750	61.650	-0.900
Qwen2-1.5B	Server	0	64.225	62.950	+1.275
Qwen2-1.5B	Server	1	63.550	61.425	+2.125
Qwen2-1.5B	Server	2	61.925	61.350	+0.575
Qwen2-7B	R4	0	79.275	78.750	+0.525
Qwen2-7B	R4	1	78.950	78.650	+0.300
Qwen2-7B	R4	2	79.150	78.600	+0.550
Qwen2-7B	R8	0	79.700	79.450	+0.250
Qwen2-7B	R8	1	79.475	79.100	+0.375
Qwen2-7B	R8	2	79.350	79.350	0.000
Qwen2-7B	Server	0	79.800	79.525	+0.275
Qwen2-7B	Server	1	79.800	79.750	+0.050
Qwen2-7B	Server	2	79.925	79.425	+0.500
CoLA	Server MCC	0	42.1725	10.6565	+31.5160
CoLA	Server MCC	1	43.0108	18.4259	+24.5849
CoLA	Server MCC	2	43.8248	16.9748	+26.8500

H Complete Communication Comparison

We sum logical tensor bytes uploaded and downloaded during all 40 rounds for Qwen2-1.5B and Qwen2-7B, including the initial adapter dispatch. Backbones are preinstalled. The totals count adapter tensor payloads, excluding serialization, transport headers, and evaluation traffic. External baselines retain their native representations and three-seed results.

FedRespro sends each client the global LoRA projected to its own rank (4 or 8) and stores residual memory at the server. Each run has 40 rounds with six participating clients per round. Traffic is the sum of the per-round tensor byte counts. These measurements use an adapter-only dispatch protocol, distinct from the main performance configurations and the separate residual channel in Appendix C.2.

Table A16: Qwen2-1.5B/7B upload and download over 40 rounds, including initialization. All values are in GiB, reported as mean $\pm$ sample SD over three seeds.

	Qwen2-1.5B			Qwen2-7B		
Method	Upload	Download	Total	Upload	Download	Total
HetLoRA	0.292 ± 0.004	0.584 ± 0.009	0.876 ± 0.013	0.655 ± 0.013	1.317 ± 0.027	1.972 ± 0.040
FlexLoRA	0.310 ± 0.003	0.612 ± 0.006	0.921 ± 0.009	0.717 ± 0.007	1.416 ± 0.014	2.134 ± 0.021
FLoRA	0.310 ± 0.003	3.930 ± 0.022	4.240 ± 0.024	0.717 ± 0.007	9.102 ± 0.050	9.819 ± 0.057
RB-LoRA	0.310 ± 0.003	0.612 ± 0.006	0.921 ± 0.009	0.717 ± 0.007	1.416 ± 0.014	2.134 ± 0.021
FSLoRA	0.310 ± 0.003	0.487 ± 0.000	0.797 ± 0.003	0.717 ± 0.007	1.128 ± 0.000	1.845 ± 0.007
FedRespro	0.310 ± 0.003	0.612 ± 0.006	0.921 ± 0.009	0.717 ± 0.007	1.416 ± 0.014	2.134 ± 0.021

FedRespro's mean total traffic is 0.921 GiB at 1.5B and 2.134 GiB at 7B. The exchanged tensors are the rank-4/8 adapters; residual memory stays on the server.

I HYPERPARAMETER SENSITIVITY PROTOCOL

We hold $\lambda_t = \lambda$ constant across rounds and vary only λ within each model size. All runs use seed 0 with the same model revision, data partitions, client schedules and ranks, local optimizer, and evaluation examples. References use fp32 precision. Scores are the equally weighted mean over eight tasks, evaluated at completed round-40 server checkpoints without parameter averaging.

Table A17: Matched λ sweeps at round 40. Scores are eight-task mean server accuracy (%) for seed 0, with data, client schedules, and other training settings fixed within each model size.

Model	λ	Accuracy
0.5B	0.5	39.15
0.5B	0.9	43.00
0.5B	1	44.30
0.5B	1.1	41.95
0.5B	2	39.08
0.5B	4	37.90
1.5B	2	61.80
1.5B	2.25	60.18
7B	1.5	77.70
7B	2	79.58

I.1 TASK-LEVEL SENSITIVITY

Figure A3 holds the model, training seed, client schedule, reference precision, and rank fixed and displays each task's change relative to $\lambda = 1$. The baseline row is zero by construction. Only completed round-40 evaluations enter this figure.

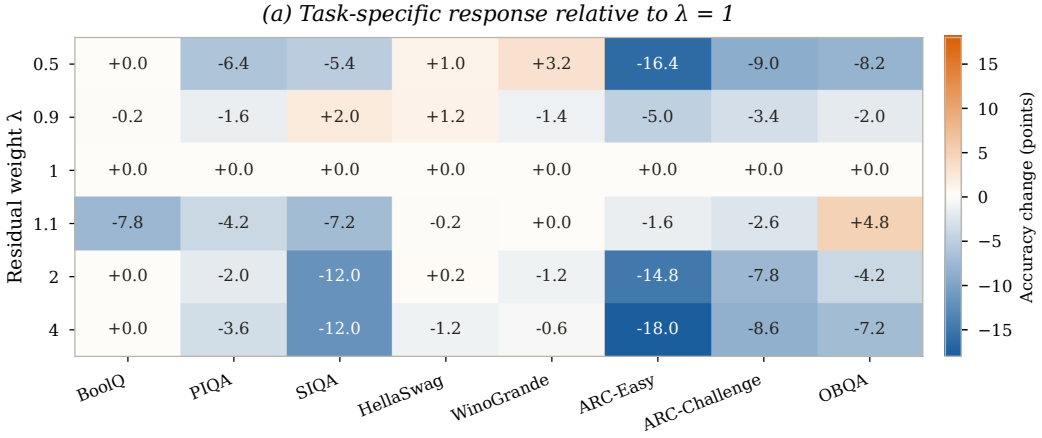

Figure A3: Task-level response to λ in Qwen2-0.5B. Each cell is the accuracy-point change from the matched $\lambda = 1$ run. All evaluations use round 40, seed 0, fp32 references, and no checkpoint averaging.